

Learning a Common Substructure of Multiple Graphical Gaussian Models

Satoshi Hara^a, Takashi Washio^a

^a*Institute of Scientific and Industrial Research (ISIR), Osaka University, Osaka, 5670047, Japan*

Abstract

Properties of data are frequently seen to vary depending on the sampled situations, which usually changes along a time evolution or owing to environmental effects. One way to analyze such data is to find invariances, or representative features kept constant over changes. The aim of this paper is to identify one such feature, namely interactions or dependencies among variables that are common across multiple datasets collected under different conditions. To that end, we propose a common substructure learning (CSSL) framework based on a graphical Gaussian model. We further present a simple learning algorithm based on the Dual Augmented Lagrangian and the Alternating Direction Method of Multipliers. We confirm the performance of CSSL over other existing techniques in finding unchanging dependency structures in multiple datasets through numerical simulations on synthetic data and through a real world application to anomaly detection in automobile sensors.

Keywords: Graphical Gaussian Model, Common Substructure, Dual Augmented Lagrangian, Alternating Direction Method of Multipliers

1. Introduction

In several real world data, such as that from the stock market (Baillie and Bollerslev, 1989), gene regulatory networks (Ahmed and Xing, 2009; Zhang et al., 2009), biomedical measurements (Varoquaux et al., 2010), or sensors in engineering systems (Idé et al., 2009), there are dynamical properties over time evolutions or due to changes in the surrounding environments. Such effects cause data to have different behaviors in each dataset collected under different conditions. One way to analyze such data is to explicitly include the change into

the model (Hamilton, 1994; Durbin et al., 2001), which usually requires detailed domain knowledge that is rarely available in most cases. Another way is to impose general and mild assumptions on the data. This kind of approach is especially common in the multi-task learning literatures (Caruana, 1997; Turlach et al., 2005), where the relationships among datasets are treated as a clue for combining multiple tasks into a single problem. The scope of the present paper is in the latter context where the relationship among datasets is the objective we want to analyze. For the purpose, we focus on invariance of the data against the underlying changes which provides partial yet important aspects of the data behaviors (von Bünau et al., 2009; Hara et al., 2012). We provide a technique for finding one of such invariance, specifically constant interactions or dependencies among variables across several different conditions. An illustrative example is an engineering system where system errors are observed as dependency anomalies in sensor values (Idé et al., 2009), which are usually caused by a fault in a subsystem. The invariance, which in this example is the remaining healthy subsystems, is captured by a steady dependency over the multiple datasets sampled before and after the error onset. Hence, we can use such information as a clue for finding erroneous subsystems.

Graphical modeling is a popular approach for analyzing dependencies in multivariate data (Lauritzen, 1996). We adopt one of the most fundamental models, a graphical Gaussian model (GGM), as the basis of our framework. A GGM is a basic model representing *linear* dependencies among *continuous* random variables, and has been widely studied owing to the simple nature, that is, the dependency structure is represented by the zero patterns in an inverse covariance matrix. Identification of such zero patterns from data was first studied by Dempster (1972) as a *Covariance Selection* where the task is formulated as the combinatorial problem of optimizing the location of zeros in a matrix. Since classical algorithms for this do not scale to high dimensional data, the scope of studies has shifted to a relaxed setting (Meinshausen and Bühlmann, 2006; Yuan and Lin, 2007; Banerjee et al., 2008), where Covariance Selection is formulated as a convex optimization problem using a ℓ_1 -regularization that induces zeros in the resulting matrix. Because of the effectiveness of the relaxed formulation, several related optimization techniques have also been studied (Friedman et al., 2008; Duchi et al., 2008a; Li and Toh, 2010; Scheinberg and Rish, 2010; Yuan, 2009; Scheinberg et al., 2010; Hsieh et al., 2011).

In our context, the objective is not to estimate the structure of a GGM

from a single dataset, but to decompose the resulting GGMs from several datasets into common and individual substructures, with the former representing the invariance we aim to detect. There are some prior studies on learning a set of GGMs from multiple datasets. Varoquaux et al. (2010) and Honorio and Samaras (2010) imported the idea of Group-Lasso (Yuan and Lin, 2006; Bach, 2008) and Multitask-Lasso (Turlach et al., 2005; Liu et al., 2009), and extended the framework of a single GGM setting. In both cases, the problem is formulated under the assumption that all matrices share the same zero patterns. Guo et al. (2011) considered a method to avoid this additional assumption, although the problem then loses convexity. Though these approaches achieved some success in improving the estimation accuracy of graphical models, this does not necessarily mean that they are suitable for finding commonness across datasets as we will see in the simulation. In the context of common substructure detection, Zhang and Wang (2010) proposed using a Fused-Lasso (Tibshirani et al., 2005) type of technique to find an invariant pattern between two datasets. As a general framework for N datasets situations, Chiquet et al. (2011) considered imposing sign coherence on the resulting structures, while Hara and Washio (2011) extended the framework of Zhang and Wang (2010) to the general situation of N datasets¹. In the opposite context where the target is dynamics rather than invariance, Zhou et al. (2010) proposed using weighted statistics to trace the evolution of a GGM. We note there are also several related studies in the binary Markov random field literatures (Guo et al., 2007; Ahmed and Xing, 2009). They also use ℓ_1 -regularization (Wainwright et al., 2007) and Fused-Lasso type techniques (Ahmed and Xing, 2009) for recovering temporal dependency structures, which are technically quite close to the ones of GGM.

The contribution of this paper is two folds. First, we introduce the novel *Common Substructure Learning* (CSSL) framework that is applicable for a general case of N datasets. Second, a sophisticated algorithm based on the Dual Augmented Lagrangian (DAL) (Tomioka et al., 2011) and the Alternating Direction Method of Multipliers (ADMM) (Gabay and Mercier, 1976; Boyd et al., 2011) is proposed. In the proposed algorithm, the inner problems for each iterative update are simple and can be solved efficiently which

¹This paper is an extension of Hara and Washio (2011) with more general settings, an efficient optimization algorithm, and exhaustive simulations on synthetic and real world datasets.

Table 1: Mathematical Notation

Notation	Description
$\ \mathbf{x}\ _p$	ℓ_p -norm of a vector $\mathbf{x} \in \mathbb{R}^d$, $\ \mathbf{x}\ _p = \left(\sum_{i=1}^d x_i ^p\right)^{\frac{1}{p}}$ for $p \in [1, \infty)$ and $\ \mathbf{x}\ _\infty = \max_{1 \leq i \leq d} x_i $
$\ A\ _p$	vectorized ℓ_p -norm of a matrix $A \in \mathbb{R}^{d \times d}$, $\ A\ _p = \ (A_{11}, A_{12}, \dots, A_{dd})^\top\ _p$
$\ A\ _S$	spectral norm of a matrix $A \in \mathbb{R}^{d \times d}$, $\ A\ _S = \max_{1 \leq i \leq d} \sigma_i(A)$ where $\sigma_i(A)$ is an i th singular value of A
$\ B\ _{1,p}$	$\ell_{1,p}$ -norm of matrices $B = \{B_i; B_i \in \mathbb{R}^{d \times d}\}_{i=1}^N$, $\ B\ _{1,p} = \sum_{j,j'=1}^d \ (B_{1,jj'}, B_{2,jj'}, \dots, B_{N,jj'})^\top\ _p$
$A \succ 0$	a matrix A is symmetric and positive definite
$\text{sgn}(a)$	sign function on a scalar a , $\text{sgn}(a) = 1$ for $a > 1$, $\text{sgn}(a) = -1$ for $a < 0$ and $\text{sgn}(a) = 0$ for $a = 0$
$\text{diag}(\mathbf{x})$	$d \times d$ matrix with $\mathbf{x} \in \mathbb{R}^d$ on its diagonal

results in fast computation. We confirm the validity of the CSSL approach through simulations on synthetic datasets and on an anomaly detection task in real-world data.

The remainder of the paper is organized as follows. In Section 2, we briefly review properties of GGMs and existing learning techniques. In Section 3, we present the proposed framework and its theoretical properties. The optimization algorithm based on DAL-ADMM is introduced in Section 4. The validity of the proposed method is presented through synthetic experiments in Section 5. In Section 6, we apply the proposed method to an anomaly detection task on *sensor error* data. Finally, we conclude the paper in Section 7.

2. Structure Learning of Graphical Gaussian Model

In this section, we review the GGM estimation problem (Meinshausen and Bühlmann, 2006; Yuan and Lin, 2007; Banerjee et al., 2008) and some prior extensions to multiple datasets (Varoquaux et al., 2010; Honorio and Samaras, 2010; Zhang and Wang, 2010).

We also summarize mathematical notations used throughout the paper in Table 1.

2.1. Graphical Gaussian Model

In multivariate analysis, covariance and correlation are commonly used as indicators for a relationship between two random variables. However, in general, a covariance between two random variables x_j and $x_{j'}$ is affected by other variables. Therefore, we need to remove such effects to estimate an essential dependency structure, which is available by searching for conditional dependency among random variables. In a general graphical model, we express these dependencies using a graph with vertices corresponding to each random variable and edges spanning random variables that are conditionally dependent.

Here, we assume that a d -dimensional random variable $\mathbf{x} = (x_1, x_2, \dots, x_d)^\top$ follows a zero mean Gaussian distribution, that is, $\mathbf{x} \sim \mathcal{N}(\mathbf{0}_d, \Lambda^{-1})$ for some symmetric and strictly positive definite matrix $\Lambda \in \mathbb{R}^{d \times d}$. We refer to a graphical model of Gaussian variables as graphical Gaussian model (GGM). Note that the zero mean assumption can be achieved without loss of generality by subtracting a sample mean from the dataset. Here, a covariance matrix is parameterized as the inverse of a *precision matrix* Λ since this is a more primitive parameter representing essential dependency among variables. A precision matrix relates to the conditional expectation as

$$\Lambda_{jj'} \propto -\mathbb{E}[x_j x_{j'} | \text{other variables}] ,$$

that is, the (j, j') th entry of Λ is proportional to the covariance between x_j and $x_{j'}$ with the remaining $d - 2$ variables fixed. With this property, the conditional independence between Gaussian random variables is expressed as zero entries of Λ :

$$\Lambda_{jj'} = 0 \Leftrightarrow x_j \perp\!\!\!\perp x_{j'} \mid \text{other variables}$$

where $\perp\!\!\!\perp$ denotes statistical independence. Because of this property, the edge patterns in a GGM correspond to the non-zero entries in a precision matrix Λ . In a GGM, two vertices have an edge between them if and only if the corresponding (j, j') th entry of Λ is non-zero. In the case that only few pairs of variables are dependent, most off-diagonal elements in Λ are zeros and the corresponding graph expression is sparse, which allows us to visually inspect the underlying relations.

2.2. Sparse Estimation of GGM

A naive way to estimate a precision matrix Λ is a maximum likelihood estimation formulated as

$$\begin{aligned}\hat{\Lambda} &= \operatorname{argmax}_{\Lambda \in \mathcal{P}} \ell(\Lambda; S), \\ \ell(\Lambda; S) &= \log \det \Lambda - \operatorname{tr} [S\Lambda].\end{aligned}\tag{1}$$

Here, $\ell(\Lambda; S)$ is a log-likelihood of a Gaussian distribution (up to a constant), S is a sample covariance matrix and \mathcal{P} is a set of symmetric positive definite matrices $\mathcal{P} = \{A \in \mathbb{R}^{d \times d}; A \succ 0\}$. The positive definiteness constraint is imposed so that the resulting Λ is a valid precision matrix. For a strictly positive definite matrix S , the solution to this problem is $\hat{\Lambda} = S^{-1}$. However, in a finite sample case, even when the true parameter is zero, that is, $\Lambda_{jj'} = 0$, its maximum likelihood estimator $\hat{\Lambda}_{jj'}$ is non-zero with probability one. In this situation, the resulting graphical model is a complete graph, which states that every pairs of variables is conditionally dependent and the underlying intrinsic relationships are masked.

The major scope of GGM studies is how to avoid this unfavorable result from a maximum likelihood estimation and infer a sparse graph structure, which is referred to as *Covariance Selection* (Dempster, 1972). In classical studies, some entries of a precision matrix Λ are fixed as zeros and the remaining non-zero entries are estimated, where the zero pattern is optimized in a combinatorial manner. However, this combinatorial problem is not feasible for high-dimensional data.

In recent studies, the use of an ℓ_1 -regularization has been shown to be practical for Covariance Selection. The first such study was conducted by Meinshausen and Bühlmann (2006). In their approach, the solution is obtained by solving the Lasso (Tibshirani, 1996). Here, let us denote d -dimensional data with n data points using an $n \times d$ matrix $X = [\mathbf{x}_1 \ \mathbf{x}_2 \ \dots \ \mathbf{x}_n]^\top$, with X_j as its j th column and $X_{\setminus j}$ as its remaining $d - 1$ columns. For each column, we solve the following Lasso:

$$\min_{\boldsymbol{\theta}} \frac{1}{2} \|X_j - X_{\setminus j} \boldsymbol{\theta}\|_2^2 + \rho \|\boldsymbol{\theta}\|_1, \tag{2}$$

where $\rho \geq 0$ is a regularization parameter. We then set zero patterns of $\boldsymbol{\theta}$ to the j th column of Λ . Meinshausen and Bühlmann (2006) have also showed the asymptotic convergence of their estimator to the true graph structure

under a proper condition. This approach was later reformulated as an ℓ_1 -regularized maximum likelihood problem (Yuan and Lin, 2007; Banerjee et al., 2008):

$$\max_{\Lambda \in \mathcal{P}} \ell(\Lambda; S) - \rho \|\Lambda\|_1 . \quad (3)$$

We refer to this problem as *Sparse Inverse Covariance Selection* (SICS) following Scheinberg et al. (2010). The resulting precision matrix of (3) has some zero entries owing to the effect of an additional ℓ_1 -regularization term. Several efficient optimization techniques are available for solving this problem. Examples include GLasso (Friedman et al., 2008), PSM (Duchi et al., 2008a), IPM (Li and Toh, 2010), SINCO (Scheinberg and Rish, 2010), ADMM (Yuan, 2009; Scheinberg et al., 2010) and QUIC (Hsieh et al., 2011).

2.3. Learning a Set of GGMs with Same Topological Patterns

The ordinary SICS problem (3) aims to learn one GGM from a single dataset. The extension of this framework to multiple datasets has been studied by Varoquaux et al. (2010) and Honorio and Samaras (2010). The task is to estimate N precision matrices $\Lambda_1, \Lambda_2, \dots, \Lambda_N$ from N datasets where the sample covariance matrices for each dataset are S_1, S_2, \dots, S_N . The objective of this multi-task extension is to improve the estimation accuracy of each GGM by incorporating the similarity among datasets. In the framework of the above studies, GGMs from each dataset are assumed to have the same topological patterns, that is, the same edge connection structures while the edge weights might be different for each GGM. They both introduced a $\ell_{1,p}$ -norm of a set of N precision matrices $\{\Lambda_i\}_{i=1}^N$

$$\|\Lambda\|_{1,p} = \sum_{j,j'=1}^d \left(\sum_{i=1}^N |\Lambda_{i,jj'}|^p \right)^{\frac{1}{p}} ,$$

as a regularization term analogous to the Group-Lasso (Yuan and Lin, 2006; Bach, 2008) and Multitask-Lasso (Turlach et al., 2005; Liu et al., 2009) with $p \in [1, \infty]$. Varoquaux et al. (2010) has considered the case $p = 2$ while Honorio and Samaras (2010) used $p = \infty$. These two choices are commonly adopted in many scenarios owing to the computational efficiency. The entire estimation problem is defined as

$$\max_{\{\Lambda_i; \Lambda_i \in \mathcal{P}\}_{i=1}^N} \sum_{i=1}^N t_i \ell(\Lambda_i; S_i) - \rho \|\Lambda\|_{1,p} , \quad (4)$$

with non-negative weights t_1, t_2, \dots, t_N . Without loss of generality, we can limit ourselves to the normalized case $\sum_{i=1}^N t_i = 1$ since the unnormalized version is just a scaled objective function for some constant. The typical choice of parameters is $t_i = \frac{n_i}{\sum_{i=1}^N n_i}$ where n_i is the number of data points in the i th dataset. We refer to problem (4) as *Multitask Sparse Inverse Covariance Selection* (MSICS) in the remainder of the paper.

Note that the MSICS problem (4) involves the ordinary SICS (3) as a special case when $p = 1$ where the $\ell_{1,1}$ -regularization term completely decouples into N individual ℓ_1 -regularizations. In the extended case for $p > 1$, the regularization term enforces the joint structure $\tilde{\Lambda}_{jj'} = \left(\sum_{i=1}^N |\Lambda_{i,jj'}|^p \right)^{\frac{1}{p}}$ to be sparse, with $\tilde{\Lambda}_{jj'} = 0$ indicating that the corresponding (j, j') th entries are zeros across all N precision matrices.

2.4. Learning Structural Changes between Two GGMs

Although taking advantage of situations with multiple datasets using the preceding techniques is useful for improving the estimation performances of the resulting GGMs, it only imposes joint zero patterns and does not indicate anything about the commonness of the non-zero entries. It is therefore not that helpful when comparing GGMs representing similar models where we expect that there may exist some common edges whose weights are close to each other. Zhang and Wang (2010) considered the two datasets case and constructed an algorithm using a Fused-Lasso type regularization (Tibshirani et al., 2005) to round these similar values to be exactly the same allowing only significantly different edges between two GGMs to be extracted. Their approach follows the ideas of Meinshausen and Bühlmann (2006) by connecting the update procedure (2) for two datasets X_1 and X_2 through a new regularization term for the variation between two parameters $\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_1$,

$$\min_{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2} \sum_{i=1}^2 \left\{ \frac{1}{2} \|X_{i,j} - X_{i,\setminus j} \boldsymbol{\theta}_i\|_2^2 + \rho \|\boldsymbol{\theta}_i\|_1 \right\} + \gamma \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_1, \quad (5)$$

where $\gamma \geq 0$ is a regularization parameter for the variation. The new term enforces the variation of some elements in two parameters to shrink to zeros. They also provided a coordinate descent-based optimization procedure for the above problem.

3. Learning Common Patterns in Multiple GGMs

The preceding work by Zhang and Wang (2010) adopted the idea of the Fused-Lasso type technique using the specific formulation of the two datasets situation. In our study, we introduce a new framework, a *Common Substructure Learning* (CSSL), for finding invariant patterns in multiple dependency structures that is applicable to the general case of N datasets.

3.1. Common Substructure Learning Problem

We first formalize what invariance we are aiming to detect in multiple dependency structures. To begin with, we assume that the number of variables in each dataset is the same, so they are all d -dimensional. Also, the identities of each variable are the same. For instance, x_1 is always a value from the same sensor while its behavior may change across datasets. We then define a common substructure for multiple GGMs as follows.

Definition 1 (Common Substructure of Multiple GGMs). *Let $\Lambda_1, \Lambda_2, \dots, \Lambda_N$ be precision matrices corresponding to each GGM. Then, the common substructure of the GGMs is expressed by an adjacency matrix $\Theta \in \mathbb{R}^{d \times d}$ defined as*

$$\Theta_{jj'} = \begin{cases} \Lambda_{1,jj'}, & \text{if } \Lambda_{1,jj'} = \Lambda_{2,jj'} = \dots = \Lambda_{N,jj'} \\ 0, & \text{otherwise} \end{cases}. \quad (6)$$

Note this is a natural extension of the invariance notion adopted in the prior work by Zhang and Wang (2010) for the case of two datasets. With an ordinal sparsity assumption for GGMs, this definition leads the precision matrices to simultaneously have sparseness and commonness. That is:

- Sparseness: $\Lambda_{i,jj'} = 0$ for some $1 \leq i \leq N$ and $1 \leq j, j' \leq d$,
- Commonness: $\Lambda_{1,jj'} = \Lambda_{2,jj'} = \dots = \Lambda_{N,jj'}$ for some $1 \leq j, j' \leq d$.

Under the above commonness, the basic idea of our framework is to parametrize each precision matrix Λ_i using two components, a common substructure Θ and an individual substructure $\Omega_i \in \mathbb{R}^{d \times d}$:

$$\Lambda_i = \Theta + \Omega_i. \quad (7)$$

Here, each individual substructure matrix Ω_i is composed of non-zero entries that are not common across the N precision matrices.

In the preceding formulation (5), some entries in the two precision matrices are shrunk to the same value owing to the effect of the term $\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_1$. In the proposed parameterization, such commonness corresponds to the case when some entries of the individual substructures are simultaneously zero, that is, $\Omega_{1,jj'} = \Omega_{2,jj'} = \dots = \Omega_{N,jj'} = 0$. Hence, the non-zero common value is expressed by a common substructure matrix Θ . These facts motivate us to regularize the individual substructures through the grouped regularization $\|\Omega\|_{1,p}$. On the other hand, we expect a common substructure Θ to be sparse so that we can interpret it easily. To that end, we adopt an ordinary ℓ_1 -regularization $\|\Theta\|_1$ and the overall problem is summarized as follows:

$$\begin{aligned} \max_{\Theta, \{\Omega_i\}_{i=1}^N} \quad & \sum_{i=1}^N t_i \ell(\Theta + \Omega_i; S_i) - \rho \|\Theta\|_1 - \gamma \|\Omega\|_{1,p} \\ \text{s.t.} \quad & \Theta + \Omega_i \in \mathcal{P} \quad (1 \leq i \leq N), \end{aligned} \quad (8)$$

with regularization parameters $\rho, \gamma \geq 0$. Since $-\ell(\Theta + \Omega_i; S_i)$, $\|\Theta\|_1$ and $\|\Omega\|_{1,p}$ are all convex, the entire formulation is again a convex optimization problem. We refer to this problem as *Common Substructure Learning* (CSSL). Note that in the above formulation, we have slightly relaxed the condition of commonness to allow $\Theta_{jj'}$ and $\Omega_{i,jj'}$ to become simultaneously non-zeros which is contrary to Definition (6). We correct this point by applying the criterion (6) to the resulting precision matrices $\hat{\Lambda}_1, \hat{\Lambda}_2, \dots, \hat{\Lambda}_N$ in the post processing stage to extract only truly common entries.

Here, we list two important properties of the CSSL problem (8), a dual problem and the bound on eigenvalues. We first present the dual problem, which plays an important role in constructing an efficient optimization algorithm in the next section.

Proposition 1 (Dual of CSSL). *The dual problem of CSSL (8) is*

$$\begin{aligned} \min_{\{W_i; W_i \in \mathcal{P}\}_{i=1}^N} \quad & - \sum_{i=1}^N t_i \log \det W_i - d, \\ \text{s.t.} \quad & \left| \sum_{i=1}^N t_i (W_{i,jj'} - S_{i,jj'}) \right| \leq \rho, \\ & \left(\sum_{i=1}^N t_i^q |W_{i,jj'} - S_{i,jj'}|^q \right)^{\frac{1}{q}} \leq \gamma \quad (1 \leq j, j' \leq d), \end{aligned} \quad (9)$$

where q is a parameter satisfying $p^{-1} + q^{-1} = 1$. The resulting matrices of the dual problem W_i^* are related to the optimal precision matrices Λ_i^* through the inverse, $\Lambda_i^* = W_i^{*-1}$.

In both the primal and dual formulations (8), (9), we enforced the positive definiteness constraints, $\Lambda_i = \Theta + \Omega_i \in \mathcal{P}$ and $W_i \in \mathcal{P}$ so that the matrices are valid precision or covariance matrices. Here, we show that they can be tightened according to the next theorem.

Theorem 1 (Bounds on Eigenvalues). *The optimal precision matrices for the CSSL (8) $\Lambda_1^*, \Lambda_2^*, \dots, \Lambda_N^*$ with $0 < \rho < N^{\frac{1}{p}}\gamma < \infty$ have bounded eigenvalues $\lambda_i^{\min} I_d \preceq \Lambda_i^* \preceq \lambda_i^{\max} I_d$, where the bounding parameters λ_i^{\min} and λ_i^{\max} are*

$$\lambda_i^{\min} = \frac{t_i}{t_i \|S_i\|_S + d\gamma}, \quad \lambda_i^{\max} = \frac{N^{\frac{1}{p}} d^2}{\rho}.$$

Using this result, we can replace the constraint $\Lambda_i \in \mathcal{P}$ with the tighter $\Lambda_i \in \tilde{\mathcal{P}}_i = \{A \in \mathbb{R}^{d \times d}; A \succeq \lambda_i^{\min} I_d\}$, and similarly $W_i \in \{A \in \mathbb{R}^{d \times d}; A \succeq \lambda_i^{\max-1} I_d\}$. Note that this update is practically important when constructing an optimization algorithm. Since the new constraint set $\tilde{\mathcal{P}}_i$ is closed, we can project points out of the constraint set onto the boundary, which is unavailable for the original open set \mathcal{P} .

3.2. Interpretations of CSSL

The proposed CSSL problem (8) can be interpreted as a generalization of an ordinary SICS problem (3) and its multi-task extension MSICS (4). In the case that $\gamma \rightarrow \infty$, the solution to the CSSL is $\Omega_1 = \Omega_2 = \dots = \Omega_N = 0_{d \times d}$, which means that all precision matrices are equal and are represented by a single matrix Θ . Such Θ is available by solving the SICS problem (3) with $S = \sum_{i=1}^N t_i S_i$. On the other hand, if $\rho \geq N^{\frac{1}{p}}\gamma$, the common substructure Θ becomes zero. This fact follows from the relationship between the ℓ_p -norms:

$$\gamma \|\Theta + \Omega_i\|_{1,p} \leq N^{\frac{1}{p}}\gamma \|\Theta\|_1 + \gamma \|\Omega_i\|_{1,p} \leq \rho \|\Theta\|_1 + \gamma \|\Omega_i\|_{1,p}.$$

Suppose that the common substructure is non-zero, that is, $\Theta \neq 0_{d \times d}$, then the above inequality means that the update $\Omega_i \leftarrow \Theta + \Omega_i$ and $\Theta \leftarrow 0_{d \times d}$ improves the objective function value (8) without changing the resulting precision matrix $\Lambda_i = \Theta + \Omega_i$, and thus the solution must be $\Theta = 0_{d \times d}$. Under

this situation, the CSSL problem (8) coincides with MSICS (4). For the proper parameters $\rho < N^{\frac{1}{p}}\gamma < \infty$, the CSSL problem (8) is the intermediate of those two problems.

The CSSL problem can also be interpreted from a distributional perspective. From the relationship between the Lagrangian expression and the constrained optimization problem, the CSSL problem (8) is equivalent to solving a set of N maximum likelihood estimation problems (1) under the additional constraints

$$\|\Theta\|_1 \leq \eta, \|\Omega\|_{1,p} \leq \eta', \quad (10)$$

for some properly chosen positive constants η, η' . Moreover, we have

$$\begin{aligned} \max_{1 \leq i < i' \leq N} \|\Omega_i - \Omega_{i'}\|_1 &\leq \max_{1 \leq i < i' \leq N} \sum_{j,j'=1}^d (|\Omega_{i,jj'}| + |\Omega_{i',jj'}|) \\ &\leq 2 \|\Omega\|_{1,\infty} \leq 2 \|\Omega\|_{1,p}, \end{aligned}$$

where the second inequality comes from the fact that exchanging the order of $\max_{1 \leq i < i' \leq N}$ and $\sum_{j,j'=1}^d$ produces the upper bound. The last inequality is an ordinary relationship between ℓ_p -norms. These relations and the fact that $\Lambda_i - \Lambda_{i'} = \Omega_i - \Omega_{i'}$ lead to the bound

$$\max_{1 \leq i < i' \leq N} \|\Lambda_i - \Lambda_{i'}\|_1 \leq 2\eta'.$$

Hence, from the result of Honorio (2011, Lemma 23) and general matrix norm rules, the left-hand side of this inequality can be interpreted as the upper bound of the KL divergence between two distributions $p_i(\mathbf{x}) = \mathcal{N}(\mathbf{0}_d, \Lambda_i^{-1})$ and $p_{i'}(\mathbf{x}) = \mathcal{N}(\mathbf{0}_d, \Lambda_{i'}^{-1})$. With these properties, we can interpret the second constraint in (10) as a constraint on the similarity among distributions:

$$\max_{1 \leq i, i' \leq N} D_{\text{KL}}(p_i(\mathbf{x}) || p_{i'}(\mathbf{x})) \leq 2\eta' \max_{1 \leq i \leq N} \|\Lambda_i^{-1}\|_{\text{S}},$$

where $D_{\text{KL}}(p_i(\mathbf{x}) || p_{i'}(\mathbf{x}))$ denotes a KL divergence between two distributions $p_i(\mathbf{x})$ and $p_{i'}(\mathbf{x})$. From Theorem 1, the optimal parameters $\Lambda_1^*, \Lambda_2^*, \dots, \Lambda_N^*$ have bounded spectral norms for a finite γ , and thus this upper bound on the KL divergence is always valid. Moreover, we can further extend this bound into the extreme case $\gamma \rightarrow \infty$ and $\eta' \rightarrow 0$. As we have discussed before, this is the case $\Omega_1 = \Omega_2 = \dots = \Omega_N = \mathbf{0}_{d \times d}$ and the problem is equivalent

to solving a single SICS problem for Θ with $S = \sum_{i=1}^N S_i$. Hence, from Banerjee et al. (2008, Theorem 1), we can see that the resulting precision matrices still have finite eigenvalues for $\rho > 0$, and the right hand side of the above inequality goes to zero. This means that the resulting distributions represented by precision matrices derived from CSSL (8) have to be similar to one another at some level and they can be even identical in the extreme case. Note that MSICS (4) is a special case of CSSL when $\Theta = 0_{d \times d}$ and thus the same upper bound holds, although there is the significant distinction that the parameter η' in MSICS (4) also affects the sparsity of the resulting precision matrices while CSSL (8) can control the sparsity through the other hyper-parameter ρ .

3.3. Connection to Additive Sparsity Models

In this section, we discuss some connections of the CSSL problem (8) to *Additive Sparsity Models* (Jalali et al., 2010; Chandrasekaran et al., 2010; Agarwal et al., 2011; Candès et al., 2011; Obozinski et al., 2011). In general additive sparsity models, the objective parameter we want to estimate is modeled as the sum of two components, as in (7). Hence, these two parameters are estimated using sparsity inducing norms such as an ℓ_1 -norm and a trace-norm. In this sense, CSSL can be seen as a specific example of additive sparsity models where we use the combination of an ℓ_1 -regularization and a group-wise regularization.

Here, we point out two close works from Jalali et al. (2010) and Chandrasekaran et al. (2010). The former considers the multi-task least squares regression problem under the combination of ℓ_1 , group-wise regularizations. Their basic idea is quite close to ours in that some regression parameters can be close to each other across datasets. They also prove the advantage of combining two regularizations over using only one theoretically and numerically. The latter study is on GGMs but with different sparsity assumptions from ours. They show that the additive sparsity model naturally appears in GGM when there are latent variables. In such a situation, the first component in the additive sparsity model corresponds to the precision matrix between observed variables while the latter component is an interaction between latent variables. This insight is also available for interpreting our model (7), that is, a common interaction among observed variables is contaminated by the effect of latent variables which are different for each dataset.

4. Optimization via DAL-ADMM

In this section, we present the optimization algorithm for solving the CSSL problem (9). Our basic approach here is to adopt the Augmented Lagrangian techniques (Hestenes, 1969; Powell, 1967). In a prior study, Tomioka et al. (2011) have shown that solving a dual problem using the Augmented Lagrangian, which is referred to as *Dual Augmented Lagrangian* (DAL), is preferable for the case when the primal loss is badly conditioned. See Tomioka et al. (2011, Table 3) and the discussion therein. This is actually the case we are faced with, as summarized in the next theorem.

Theorem 2. *The Hessian matrix of the CSSL primal loss function $\sum_{i=1}^N t_i \ell(\Theta + \Omega_i; S_i)$ is rank-deficient while the Hessian matrix of the CSSL dual loss function $-\sum_{i=1}^N t_i \log \det W_i$ is always full rank for $0 < \rho < N^{\frac{1}{p}} \gamma < \infty$.*

This fact motivates us to solve the dual problem rather than the primal problem. To that end, we construct an algorithm based on the DAL approach.

4.1. DAL-ADMM Algorithm

The basic structure of the proposed algorithm is based on the idea of DAL. However, while the original DAL requires solving the inner problem almost exactly (Tomioka et al., 2011), we take an alternative approach using ADMM (Gabay and Mercier, 1976; Boyd et al., 2011) that makes the entire procedure dramatically simple.

To begin with, we rewrite the CSSL dual problem (9) in the following equivalent form:

$$\begin{aligned} \min_{\{W_i, Y_i; W_i \in \mathcal{P}\}_{i=1}^N} & - \sum_{i=1}^N t_i \log \det W_i \\ \text{s.t. } & t_i W_i - Y_i - t_i S_i = 0 \quad (1 \leq i \leq N), \\ & \left| \sum_{i=1}^N Y_{i,jj'} \right| \leq \rho, \quad \left(\sum_{i=1}^N |Y_{i,jj'}|^q \right)^{\frac{1}{q}} \leq \gamma \quad (1 \leq j, j' \leq d). \end{aligned} \quad (11)$$

Based on this expression, we define the following Augmented Lagrangian

function:

$$\begin{aligned} \mathcal{L}_\beta(W, Y, Z) = & - \sum_{i=1}^N t_i \log \det W_i + \delta_\rho(Y) + \tilde{\delta}_\gamma^q(Y) \\ & + \text{tr} [Z^\top (TW - Y - T\Sigma)] + \frac{\beta}{2} \|TW - Y - T\Sigma\|_2^2, \end{aligned} \quad (12)$$

where β is a nonnegative parameter and Σ, W, Y and Z are the concatenated matrices $\Sigma = [S_1 \ S_2 \ \dots \ S_N]^\top$, $W = [W_1 \ W_2 \ \dots \ W_N]^\top$, $Y = [Y_1 \ Y_2 \ \dots \ Y_N]^\top$ and $Z = [Z_1 \ Z_2 \ \dots \ Z_N]^\top$, and T is as the matrix $T = \text{diag}([t_1, t_2, \dots, t_N]^\top) \otimes I_d$, where \otimes denotes the Kronecker product and I_d is the d -dimensional identity matrix. We also defined the functions $\delta_\rho(Y)$ and $\tilde{\delta}_\gamma^q(Y)$ as

$$\begin{aligned} \delta_\rho(Y) &= \begin{cases} 0, & \text{if } \left| \sum_{i=1}^N Y_{i,jj'} \right| \leq \rho \text{ for } 1 \leq j, j' \leq d, \\ \infty, & \text{otherwise} \end{cases}, \\ \tilde{\delta}_\gamma^q(Y) &= \begin{cases} 0, & \text{if } \left(\sum_{i=1}^N |Y_{i,jj'}|^q \right)^{\frac{1}{q}} \leq \gamma \text{ for } 1 \leq j, j' \leq d, \\ \infty, & \text{otherwise} \end{cases}. \end{aligned}$$

In the Augmented Lagrangian function (12), the optimal precision matrix Λ_i^* is represented by the optimal dual variable Z_i^* . This can be verified through a simple calculation. We set the derivative of the unaugmented Lagrangian $\mathcal{L}_0(W, Y, Z)$ over W_i to zeros and find that

$$W_i^{*-1} = Z_i^*,$$

which implies that $\Lambda_i^* = Z_i^*$ from Proposition 1. This follows since the solution to (11) must be the saddle point of the unaugmented Lagrangian function $\mathcal{L}_0(W, Y, Z)$.

We solve problem (11) using ADMM by iteratively applying the following three steps until convergence:

$$\begin{cases} W^{(k+1)} \in \underset{\{W_i; W_i \in \mathcal{P}\}_{i=1}^N}{\text{argmin}} \ \mathcal{L}_\beta(W, Y^{(k)}, Z^{(k)}) \\ Y^{(k+1)} \in \underset{Y}{\text{argmin}} \ \mathcal{L}_\beta(W^{(k+1)}, Y, Z^{(k)}) \\ Z^{(k+1)} = Z^{(k)} + \beta (TW^{(k+1)} - Y^{(k+1)} - T\Sigma) \end{cases}.$$

Hence, using ADMM, convergence of the dual variable Z to the optimal parameter Z^* is guaranteed as the number of iterations tends to infinity (Boyd et al., 2011, Section 3.2). This means we can find the optimal precision matrices $\Lambda_1^*, \Lambda_2^*, \dots, \Lambda_N^*$ using DAL-ADMM. In the following two subsections, we give the update procedures for W and Y .

4.2. Inner Optimization Problem: Update of W

The update of W can be factorized into N independent problems where each problem defines an update of W_i :

$$\min_{W_i \in \mathcal{P}} -t_i \log \det W_i + t_i \text{tr} \left[Z_i^{(k)\top} W_i \right] + \frac{\beta}{2} \left\| t_i W_i - Y_i^{(k)} - t_i S_i \right\|_2^2.$$

By setting the derivative over W_i to zero, we obtain

$$W_i - \left(\frac{1}{t_i} Y_i^{(k)} - \frac{1}{\beta t_i} Z_i^{(k)} + S_i \right) - \frac{1}{\beta t_i} W_i^{-1} = 0_{d \times d}.$$

Now, write the eigen-decomposition as $\frac{1}{t_i} Y_i^{(k)} - \frac{1}{\beta t_i} Z_i^{(k)} + S_i = P D P^\top$ with $D = \text{diag}(\sigma_1, \sigma_2, \dots, \sigma_d)$ and $P^\top P = P P^\top = I_d$. Then, the above matrix equation has a solution of the form $W_i = P \tilde{D} P^\top$ with $\tilde{D} = \text{diag}(\tilde{\sigma}_1, \tilde{\sigma}_2, \dots, \tilde{\sigma}_d)$. The equation for each eigenvalue is $\tilde{\sigma}_m - \sigma_m - \frac{1}{\beta t_i} \tilde{\sigma}_m^{-1} = 0$ ($1 \leq m \leq d$), which has the analytic solution

$$\tilde{\sigma}_m = \frac{\sigma_m + \sqrt{\sigma_m^2 + \frac{4}{\beta t_i}}}{2}.$$

Note the positive definiteness of W_i is automatically fulfilled since $\tilde{\sigma}_m > 0$ for $\beta > 0$.

4.3. Inner Optimization Problem: Update of Y

The update of Y is formulated as

$$\min_Y \delta_\rho(Y) + \tilde{\delta}_\gamma^q(Y) - \text{tr} \left[Z^{(k)\top} Y \right] + \frac{\beta}{2} \left\| T W^{(k+1)} - Y - T \Sigma \right\|_2^2,$$

or equivalently, the projection $Y = \text{proj}(Y_0, \mathcal{A})$ of $Y_0 = T W^{(k+1)} + \frac{1}{\beta} Z^{(k)} - T \Sigma$

onto the set $\mathcal{A} = \left\{ Y = \begin{bmatrix} Y_1 & Y_2 & \dots & Y_N \end{bmatrix}^\top; \left| \sum_{i=1}^N Y_{i,jj'} \right| \leq \rho, \left(\sum_{i=1}^N |Y_{i,jj'}|^q \right)^{\frac{1}{q}} \leq \gamma, \forall j, j' \right\}$,

where $\text{proj}(*, *)$ is a projection function defined as

$$\text{proj}(V, \mathcal{B}) = \underset{U \in \mathcal{B}}{\text{argmin}} \frac{1}{2} \|U - V\|_2^2.$$

Table 2: Solutions to problem (13) for $q = 1, 2$ and ∞ : see the corresponding appendix for further details. An operator $T_\gamma(\cdot)$ in $\mathbf{y} \in \partial\mathcal{C}_2$ for $q = \infty$ is a thresholding for each $y_{0,i}$, that is, $y_i = \text{sgn}(y_{0,i}) \min(|y_{0,i}|, \gamma)$.

	$q = 1$	$q = 2$	$q = \infty$
$\mathbf{y}_0 \in \mathcal{C}$	$\mathbf{y} = \mathbf{y}_0$		
$\mathbf{y} \in \partial\mathcal{C}_1$	$\mathbf{y} = \mathbf{y}_0 - \frac{\mathbf{1}_N^\top \mathbf{y}_0 - \rho \text{sgn}(\mathbf{1}_N^\top \mathbf{y}_0)}{N} \mathbf{1}_N$ (Appendix A.1)		
$\mathbf{y} \in \partial\mathcal{C}_2$	Continuous Quadratic Knapsack Problem (Appendix A.2)	$\mathbf{y} = \frac{\gamma}{\ \mathbf{y}_0\ _2} \mathbf{y}_0$ (Appendix A.3)	$\mathbf{y} = T_\gamma(\mathbf{y}_0)$ (Appendix A.4)
$\mathbf{y} \in \partial\mathcal{C}_3$	Continuous Quadratic Knapsack Problem (Appendix A.5)	Analytic Solution (Appendix A.6)	Continuous Quadratic Knapsack Problem (Appendix A.7)

We can further decompose this problem into $\mathcal{O}(d^2)$ problems over $\mathbf{y} = (Y_{1,jj'}, Y_{2,jj'}, \dots, Y_{N,jj'})^\top$ for each (j, j') th entry. Hence, each problem is

$$\mathbf{y} = \text{proj}(\mathbf{y}_0, \mathcal{C}) , \quad (13)$$

where \mathbf{y}_0 is an N -dimensional vector with the i th component equal to $y_{0,i} = t_i W_{i,jj'}^{(k+1)} + \frac{1}{\beta} Z_{i,jj'}^{(k)} - t_i S_{i,jj'}$, and where the constraint set is $\mathcal{C} = \{\mathbf{u} \in \mathbb{R}^N; |\mathbf{1}_N^\top \mathbf{u}| \leq \rho, \|\mathbf{u}\|_q \leq \gamma\}$ with $\mathbf{1}_N$ being an N -dimensional vector of ones.

For any $q \in [1, \infty]$, problem (13) has a trivial solution $\mathbf{y} = \mathbf{y}_0$ if $\mathbf{y}_0 \in \mathcal{C}$. In the remaining cases, that is, $|\mathbf{1}_N^\top \mathbf{y}_0| > \rho$ or $\|\mathbf{y}_0\|_q > \gamma$, the solution is on the boundary of the constraint set $\partial\mathcal{C} = \{\mathbf{u}; |\mathbf{1}_N^\top \mathbf{u}| = \rho, \|\mathbf{u}\|_q \leq \gamma\} \cap \{\mathbf{u}; |\mathbf{1}_N^\top \mathbf{u}| \leq \rho, \|\mathbf{u}\|_q = \gamma\}$ owing to the convexity of the objective function. Thus, the problem can be reduced to a search of the boundary. However, even though the constraint set \mathcal{C} is convex, it is an intersection of two sets and the shape of the boundary $\partial\mathcal{C}$ is rather complicated. Therefore, we do not search the boundary $\partial\mathcal{C}$ directly, but solve a set of simpler problems instead. The basic approach is to classify the boundary into three parts, $\partial\mathcal{C}_1 = \{\mathbf{u}; |\mathbf{1}_N^\top \mathbf{u}| = \rho, \|\mathbf{u}\|_q \neq \gamma\}$, $\partial\mathcal{C}_2 = \{\mathbf{u}; |\mathbf{1}_N^\top \mathbf{u}| \neq \rho, \|\mathbf{u}\|_q = \gamma\}$ and $\partial\mathcal{C}_3 = \{\mathbf{u}; |\mathbf{1}_N^\top \mathbf{u}| = \rho, \|\mathbf{u}\|_q = \gamma\}$. The problems we solve here are modified versions of (13), replacing the constraint with $\mathbf{y} \in \partial\mathcal{C}_m$ for each $m \in \{1, 2, 3\}$:

$$\mathbf{y} = \text{proj}(\mathbf{y}_0, \partial\mathcal{C}_m) . \quad (14)$$

Note that $\partial\mathcal{C}_1$ and $\partial\mathcal{C}_2$ involve infeasible solutions to the problem (13). For example, a point \mathbf{y} with $\|\mathbf{y}\|_q > \gamma$ is infeasible even if $\mathbf{y} \in \partial\mathcal{C}_1$, while these three regions covers the entire boundary of the constraint set $\partial\mathcal{C} \subset \cup_{m=1}^3 \partial\mathcal{C}_m$. This guarantees that we can search the entire boundary $\partial\mathcal{C}$ indirectly by searching the sets $\partial\mathcal{C}_m$ ($m = 1, 2, 3$) instead. Hence, if neither of the solutions to (14) for $\mathbf{y} \in \partial\mathcal{C}_1$ and $\mathbf{y} \in \partial\mathcal{C}_2$ are involved in \mathcal{C} , the solution to (13) is in $\partial\mathcal{C}_3$. We can take advantage of this property to construct an efficient solution procedure. We first solve problems (14) for $\mathbf{y} \in \partial\mathcal{C}_1$ and $\mathbf{y} \in \partial\mathcal{C}_2$, respectively, and if neither of solutions is in \mathcal{C} , then we solve (14) for $\mathbf{y} \in \partial\mathcal{C}_3$. In this paper, we focus on the specific cases $q = 1, 2$ and ∞ , since efficient solution procedures are available. In Table 2, we summarized the solutions to problem (13). For further details, see Appendix A.

4.4. Convergence Criteria

Although the asymptotic convergence of $Z^{(k)}$ as $k \rightarrow \infty$ is theoretically guaranteed, in practice we need to stop the iteration at some point. A major stopping criterion is the duality-gap, the difference between the primal and dual objective function values. Let $f(W)$ be the objective function in (9) and let $g(\Theta, \Omega)$ be the one in (8). Then the duality-gap at the k th iteration is defined as

$$\text{duality-gap} = f(\tilde{W}^{(k)}) - \max_{1 \leq k' \leq k} g(\tilde{\Theta}^{(k')}, \tilde{\Omega}^{(k')}),$$

where $\tilde{W}^{(k)}$, $\tilde{\Theta}^{(k)}$ and $\tilde{\Omega}^{(k)}$ denote parameters estimated in the k th step after proper projections and transformations. We need these modifications of variables since the estimators in intermediate steps are not necessarily feasible. For example, $W^{(k)}$ does not need to satisfy the constraints in (9) since they are imposed only on a variable Y in the DAL-ADMM setting (11). The projected variable $\tilde{W}^{(k)}$ is $\tilde{W}^{(k)} = T^{-1}\tilde{Y}^{(k)} + \Sigma$ where $\tilde{Y}^{(k)} = \text{proj}(Y_0^{(k)}, \mathcal{A})$ and $Y_0^{(k)} = T(W^{(k)} - \Sigma)$. The same goes for $\Lambda^{(k)} = Z^{(k)}$. An estimator $\Lambda_i^{(k)}$ is not necessarily positive definite, and thus we project them as $\tilde{\Lambda}_i^{(k)} = \text{proj}(\Lambda_i^{(k)}, \tilde{\mathcal{P}}_i)$. This projection is available in the following manner. Let $\Lambda_i^{(k)} = PDP^\top$ be an eigen-decomposition with a diagonal matrix $D = \text{diag}(\sigma_1, \sigma_2, \dots, \sigma_d)$. Then the projected matrix is $\tilde{\Lambda}_i^{(k)} = P\tilde{D}P^\top$, where each element of $\tilde{D} = \text{diag}(\tilde{\sigma}_1, \tilde{\sigma}_2, \dots, \tilde{\sigma}_d)$ is $\tilde{\sigma}_m = \max(\sigma_m, \lambda_i^{\min})$. For computing the value of $g(\tilde{\Theta}^{(k)}, \tilde{\Omega}^{(k)})$, we need to further factorize $\tilde{\Lambda}^{(k)}$ into $\tilde{\Theta}^{(k)}$

and $\tilde{\Omega}^{(k)}$. This can be computed in an element-wise manner. Let $\theta = \tilde{\Theta}_{jj'}^{(k)}$, $\Omega_{i,jj'}^{(k)} = \tilde{\Lambda}_{i,jj'}^{(k)} - \theta$ and $\boldsymbol{\lambda} = (\tilde{\Lambda}_{1,jj'}^{(k)}, \tilde{\Lambda}_{2,jj'}^{(k)}, \tilde{\Lambda}_{N,jj'}^{(k)})^\top$. Then the problem we need to solve is

$$\min_{\theta} \rho|\theta| + \gamma \|\boldsymbol{\lambda} - \theta \mathbf{1}_N\|_p .$$

For $p = 1$ and ∞ , this function is piecewise linear with breakpoints $\{0, \lambda_1, \lambda_2, \dots, \lambda_N\}$ and $\{0, \frac{\min_i \lambda_i + \max_{i'} \lambda_{i'}}{2}\}$, respectively. Hence, the optimal θ is one of these breakpoints and can be found by searching the candidates. For the case $p = 2$, the analytic solution is

$$\theta = \frac{1}{N} \left\{ \mathbf{1}_N^\top \tilde{\boldsymbol{\lambda}} - \operatorname{sgn} \left(\mathbf{1}_N^\top \tilde{\boldsymbol{\lambda}} \right) \sqrt{(\mathbf{1}_N^\top \tilde{\boldsymbol{\lambda}})^2 - N \frac{\gamma^2 (\mathbf{1}_N^\top \tilde{\boldsymbol{\lambda}})^2 - \rho^2 \|\tilde{\boldsymbol{\lambda}}\|_2^2}{\gamma^2 N - \rho^2}} \right\} .$$

Some other useful gaps are provided by Boyd et al. (2011). The *primal-gap* measures how much the equality constraints in (11) is fulfilled,

$$\text{primal-gap} = \|TW^{(k)} - Y^{(k)} - T\Sigma\|_2 ,$$

while the *dual-gap* is a degree of the feasibility condition of the solution, defined as

$$\text{dual-gap} = \beta \|T(Y^{(k+1)} - Y^{(k)})\|_2 .$$

In our simulations in Sections 5 and 6, we have evaluated both criteria. We set two threshold parameters ϵ_{gap} and ϵ_{pdgap} , and evaluated the conditions $\text{duality-gap} \leq \epsilon_{\text{gap}}$ and $\max(\text{primal-gap}, \text{dual-gap}) \leq \epsilon_{\text{pdgap}}$ in each iteration. If one of two conditions is fulfilled, we regard the iteration as converged and output the result. In the simulations in Sections 5 and 6, we set $\epsilon_{\text{gap}} = 10^{-5}d$ and $\epsilon_{\text{pdgap}} = 10^{-5}$.

4.5. Computational Complexity

In this section, we summarize the computational complexity of the proposed algorithm. In the W update step, the computational cost is dominated by the eigen-decomposition of a $d \times d$ matrix, which requires $\mathcal{O}(d^3)$ operations, so the overall complexity is $\mathcal{O}(Nd^3)$ for the update of N matrices. In the Y update step, we need a projection $\text{proj}(Y_0, \mathcal{A})$ which is divided into

$\mathcal{O}(d^2)$ subproblems. For both $q = 1$ and $q = \infty$, the most computationally expensive procedure is solving the continuous quadratic knapsack problem which requires sorting $\mathcal{O}(N)$ elements and has complexity $\mathcal{O}(N \ln N)$ ². In the case $q = 2$, the update is analytically available with $\mathcal{O}(N)$ complexity. The overall complexity for the Y update is thus $\mathcal{O}((N \ln N)d^2)$ for $q = 1, \infty$ and $\mathcal{O}(Nd^2)$ for $q = 2$. The complexity for the Z update is $\mathcal{O}(Nd^2)$. In the convergence check, we need to calculate the projection $\text{proj}(\Lambda_i^{(k)}, \tilde{\mathcal{P}}_i)$ which has $\mathcal{O}(d^3)$ complexity or $\mathcal{O}(Nd^3)$ for N matrices. We also need the projection $\text{proj}(Y_0^{(k)}, \mathcal{A})$ which is again $\mathcal{O}((N \ln N)d^2)$ for $q = 1, \infty$ and $\mathcal{O}(Nd^2)$ for $q = 2$. Summarizing the above results, we conclude that the computational complexity of one update in DAL-ADMM is $\mathcal{O}(Nd^3 + (N \ln N)d^2)$ for $q = 1, \infty$ and $\mathcal{O}(Nd^3)$ for $q = 2$. In many practical situations, the number of datasets N is in the tens, while the dimensionality of the data d can be a few hundred. In such cases, $\ln N \ll d$ holds, and the entire complexity is approximately $\mathcal{O}(Nd^3)$. We note this is the least necessary complexity. For an unregularized setting, the solution Λ_i^* is a maximum likelihood estimate S_i^{-1} , which requires $\mathcal{O}(d^3)$ complexity for a matrix inverse and $\mathcal{O}(Nd^3)$ for N matrices.

Despite the theoretical complexity, the choice of β is of practical importance since it affects the number of iterations needed until convergence. We propose using the heuristic from Boyd et al. (2011). In this heuristic, we update the value of $\beta = \beta^{(k)}$ in every steps following the next rule:

$$\beta^{(k+1)} = \begin{cases} 2\beta^{(k)}, & \text{if primal-gap} \geq 10 * \text{dual-gap} \\ 0.5\beta^{(k)}, & \text{if dual-gap} \geq 10 * \text{primal-gap} \\ \beta^{(k)}, & \text{otherwise} \end{cases}.$$

While this does not give any theoretical guarantees on its performance, it does give us a pragmatic choice of β and results in convergence with a smaller number of steps.

4.6. Heuristic Choice of Hyper-parameters

In the CSSL problem (8), the choice of hyper-parameters ρ and γ affects the resulting precision matrices. There are several approaches for choosing these, such as cross-validation (Yuan and Lin, 2007; Guo et al., 2011) or the

²See Appendix A.2, Appendix A.5, and Appendix A.7.

Bayesian information criterion (Guo et al., 2011). Apart from selection techniques, the following result gives us some insight into ρ and γ , and is helpful for analyzing the data more intensively.

Proposition 2. *Let the bivariate common substructure Θ and individual substructures Ω_i be in the forms $\Theta = \begin{bmatrix} 0 & \theta \\ \theta & 0 \end{bmatrix}$ and $\Omega_i = \begin{bmatrix} u_i & \omega_i \\ \omega_i & v_i \end{bmatrix}$, and consider the following CSSL problem with regularizations only on off-diagonal entries:*

$$\begin{aligned} \max_{\Theta, \{\Omega_i\}_{i=1}^N} \quad & \sum_{i=1}^N t_i \ell(\Theta + \Omega_i; S_i) - 2\rho|\theta| - 2\gamma\|\boldsymbol{\omega}\|_p \\ \text{s.t.} \quad & \Theta + \Omega_i \in \mathcal{P} \quad (1 \leq i \leq N), \end{aligned} \quad (15)$$

where $\boldsymbol{\omega} = (\omega_1, \omega_2, \dots, \omega_N)^\top$. Then the off-diagonal entries of the resulting precision matrices $\theta, \boldsymbol{\omega}$ have the following property:

$$\max_{1 \leq i \leq N} |r_i| \leq \gamma \text{ and } \left| \sum_{i=1}^N t_i r_i \right| \leq \rho \Rightarrow \theta = 0, \boldsymbol{\omega} = \mathbf{0}_N,$$

where r_i is the off-diagonal entry of S_i .

Although the result is specific to the bivariate case, we can use this as a guideline for choosing the hyper-parameters ρ and γ . It also shows that ρ and γ are not independent of each other, but rather they should change simultaneously proportional to $\max_{1 \leq i \leq N} |r_i|$ and $\left| \sum_{i=1}^N t_i r_i \right|$. In particular, if each matrix S_i is multiplied by some positive constant c , the above condition indicates that ρ and γ also need to be multiplied by c . Such scale invariance is maintained only by a linear model between ρ and γ . Therefore, we construct the following heuristic based on this linear model.

1. Assume that the linear relation $\left| \sum_{i=1}^N t_i S_{i,jj'} \right| = s_1 \max_{1 \leq i \leq N} |S_{i,jj'}| + s_0$ holds for all entries $1 \leq j \leq j' \leq d$ for some $s_0, s_1 \in \mathbb{R}$.
2. Estimate s_0, s_1 with least squares regression using the tuples $\left\{ \max_{1 \leq i \leq N} |S_{i,jj'}|, \left| \sum_{i=1}^N t_i S_{i,jj'} \right| \right\}$.
3. Parameterize ρ, γ as $\rho = \max(s_1\alpha + s_0, 0)$ and $\gamma = \alpha$ using a parameter α .

This procedure provides an efficient way of tuning ρ and γ simultaneously through a single parameter α .

5. Simulation

In this section, we investigate the performance of the proposed CSSL approach in finding common substructures among datasets through numerical simulations.

5.1. Generation of Synthetic Data

We first briefly summarize the data generation procedure for our simulations. For the synthetic data, we need N precision matrices with sparseness and commonness. We tackle this problem in a two-stage approach. We first generate a single sparse precision matrix, and then add some non-zero entries to make N matrices where the additional patterns are individual to each other³. After N precision matrices $\Lambda_1, \Lambda_2, \dots, \Lambda_N$ have been constructed, we generate N datasets from the corresponding Gaussian distributions $\mathcal{N}(\mathbf{0}_d, \Lambda_i^{-1})$ for $1 \leq i \leq N$.

5.2. Baseline Methods and Evaluation Measurements

In the simulation, we adopt SICS (3) and MSICS (4) as baseline methods to compare with CSSL. Since neither method is designed for finding a common substructure, we apply a heuristic to extract the substructure $\hat{\Theta}$ from the estimated precision matrices $\hat{\Lambda}_1, \hat{\Lambda}_2, \dots, \hat{\Lambda}_N$. Note that, in SICS, each $\hat{\Lambda}_i$ is estimated by solving (3) individually while the set of matrices is estimated simultaneously in MSICS (4). Following is the heuristic criterion used:

$$\hat{\Theta}_{jj'} = \begin{cases} \hat{\theta}_{jj'} , & \text{if } \max_{1 \leq i < i' \leq d} |\hat{\Lambda}_{i,jj'} - \hat{\Lambda}_{i',jj'}| \leq \epsilon \\ 0 , & \text{otherwise} \end{cases}$$

where ϵ is some given threshold. Here, to avoid selecting zero edges as parts of a common substructure, we set $\hat{\theta}_{jj'}$ to zero if $\hat{\Lambda}_{1,jj'} = \hat{\Lambda}_{2,jj'} = \dots = \hat{\Lambda}_{N,jj'} = 0$ and one otherwise. In our simulation, we select the threshold ϵ from the resulting precision matrices. Specifically, we compute variations of estimators for each entry $\left\{ \max_{1 \leq i < i' \leq N} |\hat{\Lambda}_{i,jj'} - \hat{\Lambda}_{i',jj'}| \right\}_{1 \leq j \leq j' \leq d}$, and then set ϵ as the $100\epsilon_0\%$ quantile. This corresponds to considering the lower $100\epsilon_0\%$ varied entries as common.

In our simulation, we evaluate the common substructure detection performance through precision, recall and the F-measure. While these values are

³See Appendix B for further details.

defined based on the number of true positive, false positive and false negative detections, we slightly modify these measurements. This is because finding common dependencies with higher amplitudes is much more important than finding very small dependencies which can be approximated as zero in practice. To that end, we adopt following weighted measurements, namely WTP (weighted true positive), WFP (weighted false positive), and WFN (weighted false negative),

$$\begin{aligned} \text{WTP} &= \sum_{j < j'}^d \tilde{J}_{c,jj'} \tilde{J}_{p,jj'} J_{c,jj'} \max_{1 \leq i \leq N} |\Lambda_{i,jj'}|, \\ \text{WFP} &= \sum_{j < j'}^d \tilde{J}_{c,jj'} \tilde{J}_{p,jj'} (1 - J_{c,jj'}) \max_{1 \leq i \leq N} |\Lambda_{i,jj'}|, \\ \text{WFN} &= \sum_{j < j'}^d \left\{ \tilde{J}_{c,jj'} (1 - \tilde{J}_{p,jj'}) + (1 - \tilde{J}_{c,jj'}) \right\} J_{c,jj'} \max_{1 \leq i \leq N} |\Lambda_{i,jj'}|, \end{aligned}$$

where $\tilde{J}_{c,jj'}$, $\tilde{J}_{p,jj'}$ and $J_{c,jj'}$ are defined as

$$\begin{aligned} \tilde{J}_{c,jj'} &= I \left(\max_{1 \leq i < i' \leq N} |\hat{\Lambda}_{i,jj'} - \hat{\Lambda}_{i',jj'}| < \epsilon \right), \\ \tilde{J}_{p,jj'} &= I \left(\max_{1 \leq i \leq N} |\hat{\Lambda}_{i,jj'}| > 0 \right), \\ J_{c,jj'} &= I \left(\max_{1 \leq i < i' \leq N} |\Lambda_{i,jj'} - \Lambda_{i',jj'}| = 0 \right). \end{aligned}$$

Here, $I(P)$ is an indicator function that returns 1 for a true statement P and 0 otherwise. The modified measurements in the simulation are defined using these values as

$$\begin{aligned} \text{Precision} &= \frac{\text{WTP}}{\text{WTP} + \text{WFP}}, \\ \text{Recall} &= \frac{\text{WTP}}{\text{WTP} + \text{WFN}}, \\ \text{F-measure} &= 2 \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}. \end{aligned}$$

In the simulation, we also observe whether the zero pattern in the precision matrices is properly recovered using each method. We use the following

F-measure for this evaluation, which we refer to the "F₀-measure" to distinguish it from the one above:

$$\begin{aligned}
\text{F}_0\text{-measure} &= \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}, \\
\text{TP} &= \sum_{i=1}^N \sum_{j < j'}^d I(\Lambda_{i,jj'} = 0) I(\hat{\Lambda}_{i,jj'} = 0), \\
\text{FP} &= \sum_{i=1}^N \sum_{j < j'}^d I(\Lambda_{i,jj'} \neq 0) I(\hat{\Lambda}_{i,jj'} = 0), \\
\text{FN} &= \sum_{i=1}^N \sum_{j < j'}^d I(\Lambda_{i,jj'} = 0) I(\hat{\Lambda}_{i,jj'} \neq 0).
\end{aligned}$$

5.3. Result

We conducted simulations for three cases with data dimensionality $d = 25, 50$ and 100 where the number of datasets is fixed at $N = 5$. For each case, we generate precision matrices $\Lambda_1, \Lambda_2, \dots, \Lambda_N$ to have 15% non-zero entries on average. In the simulation, we randomly generate datasets 100 times and applied each method using several different hyper-parameters, where in each run we set the number of data points in each dataset to be $5d$. For CSSL, we use the heuristic with a parameter α varying from 10^{-2} to 10^{-0} over 41 values. We also evaluate results for $\rho = \alpha$ and $\gamma = \infty$ to see the effect of γ in an extreme case. As discussed in Section 3.2, this corresponds to solving a single SICS problem with $S = \sum_{i=1}^N t_i S_i$ and setting the result to $\hat{\Lambda}_1 = \hat{\Lambda}_2 = \dots = \hat{\Lambda}_N = \hat{\Lambda}$. For SICS and MSICS, we set the value of ρ as $\rho = \alpha$. For each method, we adopt the resulting precision matrices with 15% non-zero entries among these 41 values of α . In SICS and MSICS, we also vary the thresholding parameter ϵ_0 between 0.5, 0.7 and 0.9.

We summarize the results in Table 3. From the table, we can see the clear advantage of CSSL for $p = 2$ and ∞ over the other methods. These two methods show higher F-measures, which are from their higher precision and recall. This contrasts with other methods, SICS and MSICS, which achieve high recall, but have relatively poor precision. This means that structure detected by those methods involve not only true common substructure but also many false detections. This shows the drawback of estimated precision matrices derived through SICS and MSICS, that is, their estimators tend to be highly varied even for true common entries while this is not the case for

CSSL. This phenomenon is especially significant in SICS, which can hardly find common substructures owing to its highly varied estimators. The results for MSICS under $p = \infty$ and $\epsilon_0 = 0.9$ are still better than the others, although $\epsilon_0 = 0.9$ means that 90% of estimated non-zero entries are considered common, which is too optimistic. Moreover, we can see that the improvement of the F-measure is achieved by the growth of recall by contrasting the results with $\epsilon_0 = 0.5$ and 0.9. This means that variations on the true common substructure mostly happens in between 50% and 90% of the entire variations of the estimated precision matrices, which are highly varied and can hardly be considered common. Note that despite the significant difference in the common entry detection performance, all methods achieve comparable zero pattern identification performance as shown by the F_0 -measure. This shows that finding common entries is a different problem from the ordinal graphical model selection, and that only CSSL does well at both tasks.

We note that CSSL with $p = 1$ and $\gamma = \infty$ give two extreme results. In the former setting, the resulting precision matrices achieve higher precision with lower recall, which is very conservative, while it is the opposite in the latter setting. The first result is caused by the difference of a grouped regularization $\|\Omega\|_{1,p}$ for $p = 1$ and $p > 1$. For $p = 1$, $\|\Omega\|_{1,p}$ completely decouples into ordinary ℓ_1 -regularizations and the resulting precision matrices do not necessarily have common zero entries in individual substructures. Intuitively speaking, the results for $p = 1$ have common zero entries $\Omega_{1,jj'} = \Omega_{2,jj'} = \dots = \Omega_{N,jj'} = 0$ only when it is strongly confident, which results in a very conservative performance compared with $p > 1$. On the other hand, if $\gamma = \infty$, the entire structures are considered to be common, which results in fewer false negatives and more false positives.

6. Application to Anomaly Detection

In this section, we apply CSSL to an anomaly detection problem. The task is to identify contributions of each variable to the difference between two datasets. Correlation anomalies (Idé et al., 2009), or errors on dependencies between variables, are known to be difficult to detect using existing approaches, especially with noisy data. To overcome this problem, the use of sparse precision matrices was proposed by Idé et al. (2009), since the sparse approach reasonably suppresses the pseudo-correlation among variables caused by noise and improves the detection rate. Here, we propose using CSSL. There is a clear indication that the proposed method can fur-

ther suppress the variation in the estimated matrices. In particular, we expect that dependency structures among healthy variables are estimated to be common, which reduces the risk that such variables are mis-detected and only anomalies are enhanced.

6.1. Anomaly Score

We adopt the measurement for correlation anomalies proposed by Idé et al. (2009). This score is based on the KL-divergence between two conditional distributions. Formally, let $\mathbf{x}^A, \mathbf{x}^B \in \mathbb{R}^d$ be Gaussian random variables following $\mathcal{N}(\mathbf{0}_d, \Lambda^{A-1})$ and $\mathcal{N}(\mathbf{0}_d, \Lambda^{B-1})$, respectively. We measure the degree of anomaly between their j th variables x_j^A and x_j^B using a KL-divergence between their conditional distributions $p_A(x_j^A | \mathbf{x}_{\setminus j}^A)$ and $p_B(x_j^B | \mathbf{x}_{\setminus j}^B)$, where $\mathbf{x}_{\setminus j}^A$ and $\mathbf{x}_{\setminus j}^B$ are the remaining $d - 1$ variables. To compute the score, we first divide the precision matrix Λ^A and its inverse W^A into a $(d - 1) \times (d - 1)$ dimensional matrix, a $d - 1$ dimensional vector, and a scalar,

$$\Lambda^A = \begin{bmatrix} L_{\setminus j}^A & \mathbf{l}_{\setminus j}^A \\ \mathbf{l}_{\setminus j}^A & \lambda_j^A \end{bmatrix}, \quad W^A = \Lambda^{A-1} = \begin{bmatrix} V_{\setminus j}^A & \mathbf{v}_{\setminus j}^A \\ \mathbf{v}_{\setminus j}^A & \sigma_j^A \end{bmatrix},$$

where we have rotated the rows and columns of Λ^A and W^A simultaneously so that their original j th rows and columns are located at the last rows and columns of the matrix. The matrices Λ^B and its inverse W^B are also divided in a same manner. The score is then given as

$$\begin{aligned} d_j^{AB} &= \int d\mathbf{x}_{\setminus j}^A p_A(\mathbf{x}_{\setminus j}^A) D_{\text{KL}}(p_A(x_j^A | \mathbf{x}_{\setminus j}^A) || p_B(x_j^B | \mathbf{x}_{\setminus j}^B)) \\ &= \mathbf{v}_{\setminus j}^{A\top} (\mathbf{l}_{\setminus j}^A - \mathbf{l}_{\setminus j}^B) + \frac{1}{2} \left\{ \frac{\mathbf{l}_{\setminus j}^{B\top} V_{\setminus j}^B \mathbf{l}_{\setminus j}^B}{\lambda_j^B} - \frac{\mathbf{l}_{\setminus j}^{A\top} V_{\setminus j}^A \mathbf{l}_{\setminus j}^A}{\lambda_j^A} \right\} \\ &\quad + \frac{1}{2} \left\{ \ln \frac{\lambda_j^A}{\lambda_j^B} + \sigma_j^A (\lambda_j^A - \lambda_j^B) \right\}. \end{aligned}$$

Here, the KL-divergence is averaged over the remaining $d - 1$ variables $\mathbf{x}_{\setminus j}^A$. Since the KL-divergence is not symmetric and $d_j^{AB} \neq d_j^{BA}$ holds in general, the resulting anomaly score a_j is decided as their maximum:

$$a_j = \max(d_j^{AB}, d_j^{BA}).$$

6.2. Simulation Setting

We evaluate the anomaly detection performance using *sensor error* data (Idé et al., 2009). The dataset comprised 42 sensor values collected from a real car in 79 normal states and 20 faulty states. The fault is caused by mis-wiring of the 24th and 25th sensors, resulting in correlation anomalies. Since sample covariances are rank-deficient in some datasets, we added 10^{-3} on their diagonal to avoid singularities.

For simulation, we randomly sample n_n datasets from the normal states and n_f datasets from the faulty states, and then estimate sparse precision matrices using six methods, CSSL with $p = 1, 2$ and ∞ , SICS (3), and MSICS (4) with $p = 2$ and ∞ . For CSSL, we adopt the heuristic and set $\rho = \max(s_1\alpha + s_0, 0)$ and $\gamma = \alpha$ for a given α , and for SICS and MSICS, we set $\rho = \alpha$. We test each method for 11 different values of α ranging from $10^{-1.5}$ to $10^{-0.5}$. The weight parameters t_i in CSSL and MSICS are set as $t_i = \frac{1}{2n_n}$ for normal datasets and $t_i = \frac{1}{2n_f}$ for faulty datasets to balance the effects from the two states. Since the anomaly score is designed only for a pair of datasets, we calculate anomaly scores for each of $n_n \times n_f$ pairs of datasets.

6.3. Result

We repeated the above procedure 100 times for 4 different settings, $[n_n, n_f] = [4, 1], [12, 3], [20, 5]$ and $[40, 10]$. For each run, we evaluated the detection performance of each method by drawing an ROC curve and measuring the area under the curve (AUC). In Table 4, we summarize the best median results for each method and setting. The table shows that CSSL with $p = 2, \infty$ and MSICS with $p = \infty$ achieve better detection performances than the others. In particular, CSSL with $p = 2$ and ∞ achieve $AUC = 1$ as their median performance in some cases. This means that they detect faulty sensors perfectly for more than half of the simulation. To see further differences, we plot the median anomaly scores derived from each method for $[n_n, n_f] = [20, 5]$ in Figure 1. From these graphs, we observe a clear distinction between successful methods and other methods on the significance of healthy sensors. The 22nd and 28th sensors are relatively highly enhanced in SICS and MSICS with $p = 2$, but are not in CSSL and MSICS with $p = \infty$. We conjecture that this is the major cause of performance differences. Interestingly, not only the 22nd and 28th sensors but most of the other healthy sensors also have the same tendencies. That is, CSSL and MSICS with $p = \infty$ reasonably

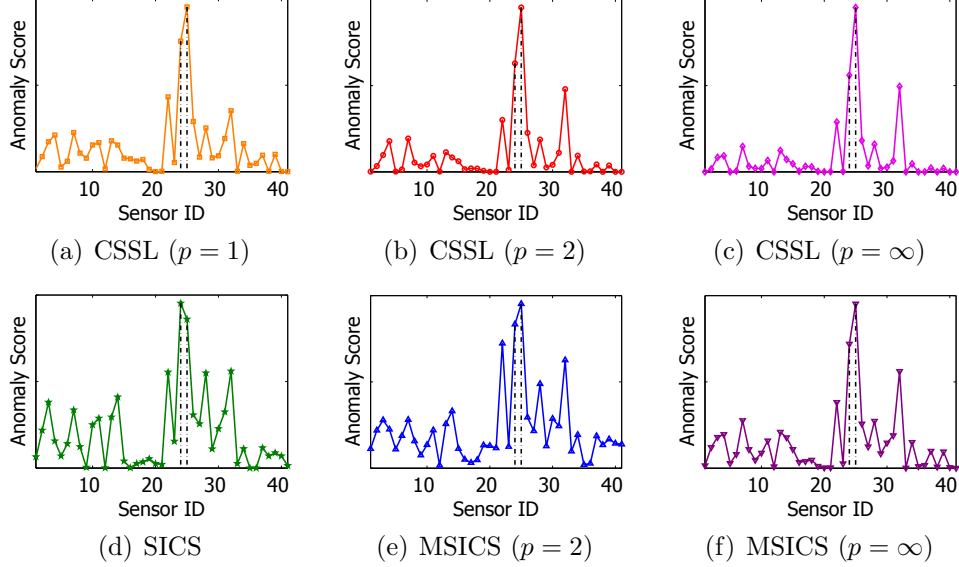


Figure 1: Median anomaly scores for each method for $[n_n, n_f] = [20, 5]$ with best AUCs. Each plot is normalized so that the maximum is the same. Dotted lines denote true faulty sensors.

suppress their significance while keeping erroneous sensors enhanced. Moreover, although the differences are subtle, we can see that CSSL with $p = 2$ and ∞ more successfully suppress the significance of sensors 1 to 21 and 33 to 42 than does MSICS with $p = \infty$. Thus, as we expected in the beginning, CSSL reduces the nuisance effects and highlights only those variables with correlation anomalies. The remaining peaks at some healthy variables are caused by the effect of the two faulty sensors since their effects may propagate to other healthy yet highly related sensors.

7. Conclusion

In this paper, we formulated the CSSL problem for multiple GGMs. We further provided a simple DAL-ADMM algorithm where each update step can be solved in a very efficient manner. Numerical results on synthetic datasets indicate the clear advantage of the CSSL approach, in that it can achieve high precision and recall at the same time, which existing GGM structure learning methods can not achieve. We also applied the proposed CSSL technique to the anomaly detection task in *sensor error* data. Through the simulation, we observed that CSSL could efficiently suppress nuisance

effects among variables in noisy sensors and successfully enhanced target faulty sensors.

Several future research topics have been indicated, including analyzing the asymptotic property of the CSSL problem (8), and extending the current formulation to the adaptive Lasso (Zou, 2006; Fan et al., 2009) type one to guarantee the *oracle property* (Zou, 2006) of the estimator. Applying the notion of commonness to more general dependency models, such as those with non-linear relations or commonness based on higher-order moment statistics, is also important.

Acknowledgments

We would like to acknowledge support for this project from the JSPS Grant-in-Aid for Scientific Research(B) #22300054. The authors would like to thank Tsuyoshi Idé and his colleagues for providing *sensor error* datasets for our simulation. We also received several helpful comments from Shohei Shimizu.

Appendix A. Solutions to (13) for $q = 1, 2$ and ∞

Here, we give detailed derivations of Table 2.

Appendix A.1. The solution is in $\partial\mathcal{C}_1$.

Problem (14) for $\mathbf{y} \in \partial\mathcal{C}_1$ is formulated as follows:

$$\min_{\mathbf{y}} \frac{1}{2} \|\mathbf{y} - \mathbf{y}_0\|_2^2 \text{ s.t. } |\mathbf{1}_N^\top \mathbf{y}| = \rho. \quad (\text{A.1})$$

Note that we have ignored the constraint $\|\mathbf{y}\|_q \neq \gamma$ because it holds for general \mathbf{y}_0 and γ with probability one. Hence, our interest is whether the solution to (A.1) satisfies $\|\mathbf{y}\|_q \leq \gamma$ or not. The additional constraint is not important in this respect.

The problem (A.1) has two possible cases as its solution, $\mathbf{1}_N^\top \mathbf{y} = \rho$ and $\mathbf{1}_N^\top \mathbf{y} = -\rho$. For each case, we can solve the problem using Lagrange multipliers:

$$\min_{\mathbf{y}} \max_{\mu} \frac{1}{2} \|\mathbf{y} - \mathbf{y}_0\|_2^2 + \mu(\mathbf{1}_N^\top \mathbf{y} - \zeta),$$

where $\zeta \in \{\rho, -\rho\}$. By setting the derivative over \mathbf{y} to zero, we get $\mathbf{y} = \mathbf{y}_0 - \mu \mathbf{1}_N$. Moreover, by substituting this result above, we derive the optimal μ as $\mu = \frac{1}{N}(\mathbf{1}_N^\top \mathbf{y}_0 - \zeta)$ and the resulting objective function value is $\frac{1}{2N}(\mathbf{1}_N^\top \mathbf{y}_0 - \zeta)^2$. The constraint $\zeta = \rho$ or $\zeta = -\rho$ is chosen so that this objective function value is minimized. Obviously, $\zeta = \rho$ is optimal for the case when $\mathbf{1}_N^\top \mathbf{y}_0 \geq 0$, while $\zeta = -\rho$ for $\mathbf{1}_N^\top \mathbf{y}_0 < 0$. Thus, the overall solution to problem (A.1) is

$$\mathbf{y} = \mathbf{y}_0 - \frac{\mathbf{1}_N^\top \mathbf{y}_0 - \rho \operatorname{sgn}(\mathbf{1}_N^\top \mathbf{y}_0)}{N} \mathbf{1}_N.$$

Appendix A.2. The solution is in $\partial\mathcal{C}_2$ for $q = 1$.

When the solution is in $\partial\mathcal{C}_2$, the problem is formulated as

$$\min_{\mathbf{y}} \frac{1}{2} \|\mathbf{y} - \mathbf{y}_0\|_2^2 \text{ s.t. } \|\mathbf{y}\|_q = \gamma. \quad (\text{A.2})$$

Here, the shape of the constraint boundary changes according to the value of q . For general $q \in [1, \infty]$, there exist several algorithms to solve this problem (Boyd and Vandenberghe, 2004; Sra, 2011). Especially, for $q = 1, 2$ and ∞ , solutions are available in a very efficient manner.

For $q = 1$, it has been shown by Honorio and Samaras (2010) that the problem is equivalent to the following *Continuous Quadratic Knapsack Problem*:

$$\min_{\mathbf{z}} \sum_{i=1}^N \frac{1}{2} (z_i - |y_{0,i}|)^2 \text{ s.t. } \mathbf{z} \geq 0, \mathbf{1}_N^\top \mathbf{z} = \gamma, \quad (\text{A.3})$$

which relates to \mathbf{y} by $y_i = \operatorname{sgn}(y_{0,i}) z_i$. Honorio and Samaras (2010) have also provided a solution technique for this problem. From the KKT condition, the solution to (A.3) is $z_i(\nu) = \max(|y_{0,i}| - \nu, 0)$ for some constant ν . Moreover, the optimal ν satisfies $\mathbf{1}_N^\top \mathbf{z}(\nu) = \gamma$. Since $\mathbf{1}_N^\top \mathbf{z}(\nu)$ is a decreasing piecewise linear function with breakpoints $|y_{0,i}|$, we can find a minimum breakpoint ν_0 that satisfies $\mathbf{1}_N^\top \mathbf{z}(\nu_0) \leq \gamma$ by sorting the N breakpoints. The optimal ν is then given as

$$\nu = \frac{\sum_{i \in \mathcal{I}_0} |y_{0,i}| - \gamma}{|\mathcal{I}_0|},$$

where $\mathcal{I}_0 = \{i; |y_{0,i}| - \nu_0 \geq 0\}$. Note that the complexity of this algorithm is $\mathcal{O}(N \log N)$ since we conduct a sorting of N values⁴.

Appendix A.3. The solution is in $\partial\mathcal{C}_2$ for $q = 2$.

The solution to problem (A.2) for $q = 2$ is analytically available. We solve the problem using Lagrange multipliers:

$$\min_{\mathbf{y}} \max_{\lambda} \frac{1}{2} \|\mathbf{y} - \mathbf{y}_0\|_2^2 + \frac{\lambda}{2} (\|\mathbf{y}\|_2^2 - \gamma^2).$$

By setting the derivative over \mathbf{y} to zero, we get $\mathbf{y} = \frac{1}{1+\lambda} \mathbf{y}_0$. Moreover, from the constraint $\|\mathbf{y}\|_2 = \gamma$, the solution is

$$\mathbf{y} = \frac{\gamma}{\|\mathbf{y}_0\|_2} \mathbf{y}_0.$$

Appendix A.4. The solution is in $\partial\mathcal{C}_2$ for $q = \infty$.

The solution of (A.2) for the case $q = \infty$ is much simpler. The problem is just a box-constrained least squares, with solution

$$y_i = \begin{cases} \gamma & (\text{if } y_{0,i} > \gamma) \\ y_{0,i} & (\text{if } |y_{0,i}| \leq \gamma) \\ -\gamma & (\text{if } y_{0,i} < -\gamma) \end{cases}$$

which is equivalent to $y_i = \text{sgn}(y_{0,i}) \min(|y_{0,i}|, \gamma)$.

Appendix A.5. The solution is in $\partial\mathcal{C}_3$ for $q = 1$.

We provide the solution procedure for (14) for $\mathbf{y} \in \partial\mathcal{C}_3$ and $q = 1$ based on the next theorem.

Theorem 3. *Let $\tilde{\mathbf{y}}$ be the solution to problem (14) for $\mathbf{y} \in \partial\mathcal{C}_1$, and suppose it is infeasible in the original problem (13). Then, the solution to (14) for $\mathbf{y} \in \partial\mathcal{C}_3$ has same signs as $\tilde{\mathbf{y}}$, that is, $\tilde{y}_i y_i \geq 0$ for $1 \leq i \leq N$.*

⁴We can further reduce this to expected linear time complexity by introducing a randomized algorithm (Duchi et al., 2008b).

From this result, we can factorize the variable indices into two parts, $\mathcal{I}_+ = \{i; \tilde{y}_i \geq 0\}$ and $\mathcal{I}_- = \{i; \tilde{y}_i < 0\}$. Using this factorization, the objective function is expressed as $\frac{1}{2} \sum_{i \in \mathcal{I}_+} (y_i - y_{0,i})^2 + \frac{1}{2} \sum_{i \in \mathcal{I}_-} (y_i - y_{0,i})^2$. The equality constraints can also be expressed as $\sum_{i \in \mathcal{I}_+} y_i + \sum_{i \in \mathcal{I}_-} y_i = \zeta$, with $\zeta \in \{\rho, -\rho\}$ and $\sum_{i \in \mathcal{I}_+} y_i - \sum_{i \in \mathcal{I}_-} y_i = \gamma$. From these expressions, we derive two independent problems:

$$\begin{aligned} \min_{\mathbf{y}^+} \frac{1}{2} \sum_{i \in \mathcal{I}_+} (y_i^+ - y_{0,i})^2 \quad \text{s.t.} \quad \mathbf{y}^+ \geq 0, \quad \sum_{i \in \mathcal{I}_+} y_i^+ &= \frac{\gamma + \zeta}{2}, \\ \min_{\mathbf{y}^-} \frac{1}{2} \sum_{i \in \mathcal{I}_-} (y_i^- + y_{0,i})^2 \quad \text{s.t.} \quad \mathbf{y}^- \geq 0, \quad \sum_{i \in \mathcal{I}_-} y_i^- &= \frac{\gamma - \zeta}{2}. \end{aligned}$$

The solutions to these problems relate to \mathbf{y} in that $y_i = y_i^+$ for $i \in \mathcal{I}_+$ and $y_i = -y_i^-$ for $i \in \mathcal{I}_-$. These problems are continuous quadratic knapsack problems and the solution can be found by using the same algorithm as in problem (A.3). We derive the final solution by solving these problems for the two cases $\zeta = \rho$ and $\zeta = -\rho$, and choosing the one with the smaller objective function value in (14).

Appendix A.6. The solution is in $\partial\mathcal{C}_3$ for $q = 2$.

The solution in the case $\mathbf{y} \in \partial\mathcal{C}_3$ and $q = 2$ is analytically available. We use Lagrange multipliers:

$$\min_{\mathbf{y}} \max_{\mu, \lambda} \frac{1}{2} \|\mathbf{y} - \mathbf{y}_0\|_2^2 + \mu(\mathbf{1}_N^\top \mathbf{y} - \zeta) + \frac{\lambda}{2} (\|\mathbf{y}\|_2^2 - \gamma^2),$$

where $\zeta \in \{\rho, -\rho\}$. By setting the derivative over \mathbf{y} to zero, we get $\mathbf{y} = \frac{1}{1+\lambda}(\mathbf{y}_0 - \mu \mathbf{1}_N)$. If $\rho = 0$, we have $\mu = \frac{\mathbf{1}_N^\top \mathbf{y}_0}{N}$ from the constraint $\mathbf{1}_N^\top \mathbf{y} = 0$. Hence, from $\|\mathbf{y}\|_2 = \gamma$, we get the optimal \mathbf{y} as

$$\mathbf{y} = \frac{\gamma}{\|\mathbf{y}_0 - \mu \mathbf{1}_N\|_2} (\mathbf{y} - \mu \mathbf{1}_N).$$

For the case $\rho > 0$, we have $\frac{1}{1+\lambda} = \frac{\zeta}{\mathbf{1}_N^\top \mathbf{y}_{0-N\mu}}$ from the constraint $\mathbf{1}_N^\top \mathbf{y} = \zeta$. Hence, we have a quadratic equation in μ from the constraint $\|\mathbf{y}\|_2^2 = \gamma^2$:

$$\rho^2 \|\mathbf{v} - \mu \mathbf{1}_N\|_2^2 = \gamma^2 (\mathbf{1}_N^\top \mathbf{v} - N\mu)^2.$$

Solving this equation gives the optimal \mathbf{y} as

$$\mathbf{y} = \frac{\zeta}{\mathbf{1}_N^\top \mathbf{y}_0 - N\mu} (\mathbf{y}_0 - \mu \mathbf{1}_N), \quad \mu = \frac{1}{N} \{ \mathbf{1}_N^\top \mathbf{y}_0 \pm \sqrt{\tau} \},$$

where $\tau = (\mathbf{1}_N^\top \mathbf{y}_0)^2 - N \frac{\gamma^2 (\mathbf{1}_N^\top \mathbf{y}_0)^2 - \rho^2 \|\mathbf{y}_0\|_2^2}{\gamma^2 N - \rho^2}$. By substituting this result into $\|\mathbf{y} - \mathbf{y}_0\|_2^2$, we have

$$\|\mathbf{y} - \mathbf{y}_0\|_2^2 = \frac{1}{N} (\zeta - \mathbf{1}_N^\top \mathbf{y}_0)^2 + \frac{N \|\mathbf{y}_0\|_2^2 - (\mathbf{1}_N^\top \mathbf{y}_0)^2}{N\tau} (\zeta \pm \sqrt{\tau})^2.$$

Since $N \|\mathbf{y}_0\|_2^2 - (\mathbf{1}_N^\top \mathbf{y}_0)^2 \geq 0$, the minimum of this value is achieved by choosing ζ and a sign in μ as $\zeta = \text{sgn}(\mathbf{1}_N^\top \mathbf{y}_0) \rho$ and $-\text{sgn}(\mathbf{1}_N^\top \mathbf{y}_0)$. Thus, the overall result is

$$\begin{aligned} \mathbf{y} &= \text{sgn}(\mathbf{1}_N^\top \mathbf{y}_0) \frac{\rho}{\mathbf{1}_N^\top \mathbf{y}_0 - N\mu} (\mathbf{y}_0 - \mu \mathbf{1}_N), \\ \mu &= \frac{1}{N} \{ \mathbf{1}_N^\top \mathbf{y}_0 - \text{sgn}(\mathbf{1}_N^\top \mathbf{y}_0) \sqrt{\tau} \}. \end{aligned}$$

Appendix A.7. The solution is in $\partial\mathcal{C}_3$ for $q = \infty$.

The solution for (14) with $\mathbf{y} \in \partial\mathcal{C}_3$ and $q = \infty$ has two possible cases, $\mathbf{1}_N^\top \mathbf{y} = \rho$ and $\mathbf{1}_N^\top \mathbf{y} = -\rho$, where for each case the problem is:

$$\min_{\mathbf{y}} \sum_{i=1}^N \frac{1}{2} (y_i - y_{0,i})^2 \text{ s.t. } \mathbf{1}_N^\top \mathbf{y} = \zeta, \quad -\gamma \mathbf{1}_N \leq \mathbf{y} \leq \gamma \mathbf{1}_N, \quad (\text{A.4})$$

with $\zeta \in \{\rho, -\rho\}$. Here, the constraint $\|\mathbf{y}\|_\infty = \gamma$ is relaxed to $\|\mathbf{y}\|_\infty \leq \gamma$. However, if the solution to (A.4) satisfies $\|\mathbf{y}\|_\infty < \rho$, it has already been found as a solution to (14) for $\mathbf{y} \in \partial\mathcal{C}_1$ and therefore this relaxation does not affect the overall procedure.

Since problem (A.4) is a variant of the continuous quadratic knapsack problem, a similar strategy to (A.3) is available. From the KKT condition, the solution to (A.4) is of the form $y_i(\nu) = \text{sgn}(y_{0,i} - \nu) \min(|y_{0,i} - \nu|, \gamma)$ for some constant ν . Moreover, the optimal ν satisfies $\mathbf{1}_N^\top \mathbf{y}(\nu) = \zeta$. Since $\mathbf{1}_N^\top \mathbf{y}(\nu)$ is a decreasing piecewise linear function with breakpoints $\{y_{0,i} - \gamma, y_{0,i} + \gamma\}_{i=1}^N$, we can find a minimum breakpoint ν_0 that satisfies $\mathbf{1}_N^\top \mathbf{y}(\nu_0) \leq \zeta$ by sorting the $2N$ breakpoints. The optimal ν is then

$$\nu = \begin{cases} \frac{\sum_{i \in \mathcal{I}_2} y_{0,i} + \gamma(|\mathcal{I}_1| - |\mathcal{I}_3|) - \zeta}{|\mathcal{I}_2|} & (\text{if } \mathcal{I}_2 \neq \emptyset) \\ \nu_0 & (\text{if } \mathcal{I}_2 = \emptyset) \end{cases}$$

where $\mathcal{I}_1 = \{i; y_{0,i} - \nu_0 \geq \gamma\}$, $\mathcal{I}_2 = \{i; -\gamma \leq y_{0,i} - \nu_0 < \gamma\}$ and $\mathcal{I}_3 = \{i; y_{0,i} - \nu_0 < -\gamma\}$.

Appendix B. Generation of Synthetic Precision Matrices

Here, we present the detailed procedure used to generate the sparse precision matrices with a common substructure in Section 5. The procedure is composed of two sequential steps. We first generate a single precision matrix, which is the common substructure in the resulting N matrices. Then, we add some non-zero entries to get a matrix Λ_i . This additional pattern is chosen to be unique for each matrix so that the resultant matrices $\Lambda_1, \Lambda_2, \dots, \Lambda_N$ satisfy the additive model assumption (7). In the following two subsections, we explain the above steps.

Appendix B.1. Generation of a Sparse Precision Matrix

In several previous studies, synthetic sparse precision matrices are generated in a naive manner, that is, just adding a properly scaled identity matrix to a sparse symmetric matrix so that the resulting matrix is sparse and positive definite (Banerjee et al., 2008; Wang et al., 2009; Li and Toh, 2010). In our simulations, we take a different approach to generating a sparse precision matrix for compatibility with the next step.

Our approach is based on an eigen-decomposition $\Lambda = VDV^\top$, where D is a matrix with eigenvalues on its diagonal and V is an orthonormal matrix such that $V^\top V = VV^\top = I_d$. Here, we use the fact that Λ is sparse if V is sufficiently sparse and the problem can be reduced to generating a sparse orthonormal matrix V . This can be done easily by applying a Givens rotation (Golub and Van Loan, 1996) to an identity matrix I_d . Formally, we let $V^{(0)} = I_d$ and apply the following procedure repeatedly until the desired sparsity is achieved.

1. Randomly pick two indices j, j' from $\{1, 2, \dots, d\}$.
2. Randomly generate θ from a uniform distribution $U([0, 2\pi])$.
3. Update the (j, j) , (j, j') , (j', j) and (j', j') th entries of $V^{(k)}$ as

$$\begin{bmatrix} V_{jj}^{(k+1)} & V_{jj'}^{(k+1)} \\ V_{j'j}^{(k+1)} & V_{j'j'}^{(k+1)} \end{bmatrix} \leftarrow \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix} \begin{bmatrix} V_{jj}^{(k)} & V_{jj'}^{(k)} \\ V_{j'j}^{(k)} & V_{j'j'}^{(k)} \end{bmatrix}.$$

4. Keep the remaining entries $V_{j_0 j'_0}^{(k+1)} \leftarrow V_{j_0 j'_0}^{(k)}$ for $(j_0, j'_0) \notin \{(j, j), (j, j'), (j', j), (j', j')\}$.

In our simulations, we generated each eigenvalue from a uniform distribution $U([0, 1])$.

Appendix B.2. Generation of Sparse Precision Matrices with a Common Substructure

Here, we turn to imposing commonness on the resulting precision matrices. To begin with, we generate small sparse precision matrices $\Psi_1, \Psi_2, \dots, \Psi_a$ in the preceding manner and construct a sparse block-diagonal precision matrix $\Lambda_0 = \text{block-diag}(\Psi_1, \Psi_2, \dots, \Psi_a)$. We then add some non-zero entries to Λ_0 and generate N precision matrices $\Lambda_1, \Lambda_2, \dots, \Lambda_N$. At this stage, we keep the original non-zero entries Λ_0 unchanged so they form a common substructure at the end. Note that the addition of non-zero entries can not be done randomly since this might destroy the positive definiteness.

We describe the procedure for the case $a = 2$. Let the eigen-decompositions of Ψ_1 and Ψ_2 be $\Psi_1 = V_1 D_1 V_1^\top$ and $\Psi_2 = V_2 D_2 V_2^\top$. Note that V_1 and V_2 are sparse since they are generated to be so. Now, let matrix Λ_i be of the form $\Lambda_i = \begin{bmatrix} \Psi_1 & \Phi_i \\ \Phi_i^\top & \Psi_2 \end{bmatrix}$. The objective is to generate a sparse non-zero matrix Φ_i while keeping the positive definiteness of Λ_i . This corresponds to keeping a determinant of Λ_i positive. Here, we choose Φ_i of the form $\Phi_i = \tilde{V}_1^b \Xi_i \tilde{V}_2^{b\top}$ where Ξ_i is a $b \times b$ diagonal matrix and \tilde{V}_1^b and \tilde{V}_2^b are matrices composed of b columns in V_1 and V_2 , respectively. Specifically, we let $V_1 = [\mathbf{v}_{1,1} \ \mathbf{v}_{1,2} \ \dots \ \mathbf{v}_{1,d_1}]$ and $V_2 = [\mathbf{v}_{2,1} \ \mathbf{v}_{2,2} \ \dots \ \mathbf{v}_{2,d_2}]$, where d_1 and d_2 denote the dimensionality of each matrix. Then \tilde{V}_1^b and \tilde{V}_2^b are $\tilde{V}_1^b = [\mathbf{v}_{1,\pi_{1,1}} \ \mathbf{v}_{1,\pi_{1,2}} \ \dots \ \mathbf{v}_{1,\pi_{1,b}}]$ and $\tilde{V}_2^b = [\mathbf{v}_{2,\pi_{2,1}} \ \mathbf{v}_{2,\pi_{2,2}} \ \dots \ \mathbf{v}_{2,\pi_{2,b}}]$, respectively, for some index sets $\{\pi_{1,1}, \pi_{1,2}, \dots, \pi_{1,b}\} \subseteq \{1, 2, \dots, d_1\}$, $\{\pi_{2,1}, \pi_{2,2}, \dots, \pi_{2,b}\} \subseteq \{1, 2, \dots, d_2\}$. Then, from a general matrix property, we can express the determinant of Λ_i as

$$\begin{aligned} \det \Lambda_i &= \det (\Psi_1 - \Phi_i \Psi_2^{-1} \Phi_i^\top) \\ &= \det (D_1 - V_1^\top \Phi_i V_2 D_2^{-1} V_2^\top \Phi_i^\top V_1) \\ &= \prod_{m=1}^b \left(\sigma_{1,\pi_{1,m}} - \frac{\xi_{i,m}^2}{\sigma_{2,\pi_{2,m}}} \right), \end{aligned}$$

where $D_1 = \text{diag}(\sigma_{1,1}, \sigma_{1,2}, \dots, \sigma_{1,d_1})$, $D_2 = \text{diag}(\sigma_{2,1}, \sigma_{2,2}, \dots, \sigma_{2,d_2})$ and $\Xi_i = \text{diag}(\xi_{i,1}, \xi_{i,2}, \dots, \xi_{i,b})$. Hence, the positive definiteness of Λ_i is guaranteed if $\xi_{i,m}^2 < \sigma_{1,\pi_{1,m}} \sigma_{2,\pi_{2,m}}$ is satisfied for $1 \leq m \leq b$. Moreover, this inequality

provides us a guideline on choosing index sets. Since we want non-zero entries of Φ_i to be larger, which can be achieved by larger $|\xi_{i,m}|$, we choose index sets so that $\sigma_{1,\pi_{1,m}}\sigma_{2,\pi_{2,m}}$ large. This corresponds to choosing leading eigenvalues and eigenvectors of Ψ_1 and Ψ_2 . In our simulations, we pick $b = 2$ indices at random from those with eigenvalues in the top $1/3$. We also generate $\xi_{i,m}$ as $\xi_{i,m} = \xi_{0,i,m}\sqrt{\sigma_{1,\pi_{1,m}}\sigma_{2,\pi_{2,m}}}$, where $\xi_{0,i,m}$ follows a uniform distribution $U([-0.8, -0.5] \cup [0.5, 0.8])$.

For general $a > 2$ cases, we first construct a matrix $\Lambda_i^{(1)}$ from Ψ_1 and Ψ_2 . We then iteratively apply the above procedure to generate $\Lambda_i^{(2)}$ from $\Lambda_i^{(1)}$ and Ψ_3 , $\Lambda_i^{(3)}$ from $\Lambda_i^{(2)}$ and Ψ_4 , until $\Lambda_i = \Lambda_i^{(a-1)}$ is derived. In the simulations in Section 5, we set the number of modules to $a = 2$ for $d = 25$, $a = 3$ for $d = 50$ and $a = 4$ for $d = 100$.

Appendix C. Proof of Theorems

Appendix C.1. Proof of Proposition 1

Let E and F_i be non-negative $d \times d$ matrices satisfying $-E_{jj'} \leq \Theta_{jj'} \leq E_{jj'}$ and $-F_{i,jj'} \leq \Omega_{i,jj'} \leq F_{i,jj'}$, respectively, for all $1 \leq i \leq N$ and $1 \leq j, j' \leq d$. Then, using Lagrange multipliers Γ, Γ_0 , and $\{\Delta_i, \Delta_{0,i}\}_{i=1}^N$, the CSSL problem (8) is expressed as

$$\begin{aligned} & \max_{\Theta, E, \{\Omega_i, F_i\}_{i=1}^N} \min_{\Gamma, \Gamma_0, \{\Delta_i, \Delta_{0,i}\}_{i=1}^N} \sum_{i=1}^N t_i \{ \log \det(\Theta + \Omega_i) - \text{tr}[S_i(\Theta + \Omega_i)] \} \\ & - \sum_{j,j'=1}^d \left\{ \rho E_{jj'} + \gamma \left(\sum_{i=1}^N F_{i,jj'}^p \right)^{\frac{1}{p}} \right\} \\ & - \text{tr}[\Gamma\Theta] + \text{tr}[\text{abs}(\Gamma)E] + \text{tr}[\Gamma_0 E] \\ & - \sum_{i=1}^N \{ \text{tr}[\Delta_i \Omega_i] - \text{tr}[\text{abs}(\Delta_i)F_i] - \text{tr}[\Delta_{0,i} F_i] \} \\ & \text{s.t. } \Gamma_{0,jj'} \geq 0, \Delta_{0,i,jj'} \geq 0 \quad (1 \leq i \leq N, 1 \leq j, j' \leq d). \end{aligned}$$

By changing the order of maximization and minimization above, we derive the dual problem. Now, we optimize each variable Θ, E, Ω_i and F_i by setting

each derivative to zero:

$$\begin{aligned}
& \sum_{i=1}^N t_i \{(\Theta + \Omega_i)^{-1} - S_i\} - \Gamma = 0_{d \times d} , \\
& -\rho \mathbf{1}_d \mathbf{1}_d^\top + \text{abs}(\Gamma) + \Gamma_0 = 0_{d \times d} , \\
& t_i \{(\Theta + \Omega_i)^{-1} - S_i\} - \Delta_i = 0_{d \times d} \quad (1 \leq i \leq N) , \\
& -\gamma \left(\sum_{i=1}^N F_{i,jj'}^p \right)^{\frac{1-p}{p}} F_{i,jj'} + |\Delta_{i,jj'}| + \Delta_{0,i,jj'} = 0 \\
& \quad (1 \leq i \leq N, 1 \leq j, j' \leq d) .
\end{aligned}$$

As a result of these equations, we get

$$\begin{aligned}
& \Delta_i = t_i \{(\Theta + \Omega_i)^{-1} - S_i\} , \\
& \left| \sum_{i=1}^N \Delta_{i,jj'} \right| \leq \rho \quad (1 \leq j, j' \leq d) , \\
& \left(\sum_{i=1}^N |\Delta_{i,jj'}|^q \right)^{\frac{1}{q}} \leq \gamma \quad (1 \leq j, j' \leq d) .
\end{aligned}$$

and so the dual problem is given by (9) where we set $W_i = (\Theta + \Omega_i)^{-1} = \frac{1}{t_i} \Delta_i + S_i$. \square

Appendix C.2. Proof of Theorem 1

We first prove the lower-bound. Let $W_i = \frac{1}{t_i} \Delta_i + S_i$ in the dual problem (9). Then we have $\left| \sum_{i=1}^N \Delta_{i,jj'} \right| \leq \rho$ and $\left(\sum_{i=1}^N |\Delta_{i,jj'}|^q \right)^{\frac{1}{q}} \leq \gamma$, and hence

$$\begin{aligned}
\left\| \frac{1}{t_i} \Delta_i + S_i \right\|_S & \leq \frac{1}{t_i} \|\Delta_i\|_S + \|S_i\|_S \\
& \leq \frac{d}{t_i} \max_{j,j'} |\Delta_{i,jj'}| + \|S_i\|_S \\
& \leq \frac{d}{t_i} \max_i \max_{j,j'} |\Delta_{i,jj'}| + \|S_i\|_S \\
& \leq \frac{d\gamma}{t_i} + \|S_i\|_S ,
\end{aligned}$$

where the last inequality comes from the general relationship between ℓ_p -norms $\max_i |\Delta_{i,jj'}| \leq \left(\sum_{i=1}^N |\Delta_{i,jj'}|^q \right)^{\frac{1}{q}}$. Since $W_i^* = \frac{1}{t_i} \Delta_i^* + S_i = \Lambda_i^{*-1}$ holds at the optimum, we have the lower-bound.

We now turn to proving the upper-bound. From strong duality, the duality-gap is zero at the optimal solution to the primal and the dual problems (8), (9), and we have

$$\rho \|\Theta^*\|_1 + \gamma \|\Omega^*\|_{1,p} = d - \sum_{i=1}^N t_i \text{tr} [S_i(\Theta^* + \Omega_i^*)] .$$

Moreover, from $0 < \rho < N^{\frac{1}{p}} \gamma < \infty$, $\text{tr} [S_i(\Theta^* + \Omega_i^*)] \geq 0$ and the general ℓ_p -norm rule $\left(\sum_{i=1}^N |\Omega_{i,jj'}^*|^p \right)^{\frac{1}{p}} \geq \max_i |\Omega_{i,jj'}^*|$,

$$\|\Theta^*\|_1 + N^{-\frac{1}{p}} \|\Omega^*\|_{1,\infty} \leq \frac{d}{\rho}$$

holds. Since $N^{\frac{1}{p}} \geq 1$ for $p \geq 1$, we get

$$\|\Theta^*\|_1 + \|\Omega^*\|_{1,\infty} \leq \frac{N^{\frac{1}{p}} d}{\rho} .$$

We use this inequality to derive the upper-bound. From the definition, the precision matrix factorizes as $\Lambda_i^* = \Theta^* + \Omega_i^*$, and hence we have

$$\begin{aligned} \|\Lambda_i^*\|_S &\leq \|\Theta^*\|_S + \|\Omega_i^*\|_S \\ &\leq \|\Theta^*\|_S + d \max_{j,j'} |\Omega_{i,jj'}^*| \\ &\leq \|\Theta^*\|_S + d \max_i \max_{j,j'} |\Omega_{i,jj'}^*| \\ &\leq \|\Theta^*\|_S + d \|\Omega^*\|_{1,\infty} \\ &\leq d \left(\|\Theta^*\|_S + \|\Omega^*\|_{1,\infty} \right) \\ &\leq d \left(\|\Theta^*\|_1 + \|\Omega^*\|_{1,\infty} \right) \\ &\leq \frac{N^{\frac{1}{p}} d^2}{\rho} \end{aligned}$$

Here, we have used the relationship $\|\Theta^*\|_S \leq \|\Theta^*\|_2 \leq \|\Theta^*\|_1$. □

Appendix C.3. Proof of Theorem 2

The Hessian matrix of the CSSL primal loss $\sum_{i=1}^N t_i \ell(\Theta + \Omega_i; S_i)$ is given by

$$\mathcal{H}_{\text{primal}} = - \begin{bmatrix} \sum_{i=1}^N t_i K_i & t_1 K_1 & t_2 K_2 & \dots & t_N K_N \\ t_1 K_1 & t_1 K_1 & 0_{d^2 \times d^2} & \dots & 0_{d^2 \times d^2} \\ t_2 K_2 & 0_{d^2 \times d^2} & t_2 K_2 & & \vdots \\ \vdots & \vdots & & \ddots & 0_{d^2 \times d^2} \\ t_N K_N & 0_{d^2 \times d^2} & \dots & 0_{d^2 \times d^2} & t_N K_N \end{bmatrix},$$

where $K_i = (\Theta + \Omega_i)^{-1} \otimes (\Theta + \Omega_i)^{-1}$. It is easy to verify that $\mathbf{1}_{N+1} \otimes I_d$ spans a null space of $\mathcal{H}_{\text{primal}}$ and thus $\mathcal{H}_{\text{primal}}$ is always rank-deficient.

On the other hand, the matrix of the CSSL dual loss $-\sum_{i=1}^N t_i \log \det W_i$ is the block-diagonal matrix

$$\mathcal{H}_{\text{dual}} = \text{block-diag}(t_1 \tilde{K}_1, t_2 \tilde{K}_2, \dots, t_N \tilde{K}_N),$$

where $\tilde{K}_i = W_i^{-1} \otimes W_i^{-1}$. From Theorem 1, we know that the CSSL solution has bounded eigenvalues and thus the above Hessian matrix is always strictly positive definite for any feasible W_i . \square

Appendix C.4. Proof of the Proposition 2

Let S_i be the covariance matrix $S_i = \begin{bmatrix} a_i & r_i \\ r_i & b_i \end{bmatrix}$. Then we have an upper-bound for (15) of

$$\begin{aligned} & \sum_{i=1}^N t_i \{ \log(u_i v_i - (\theta + \omega_i)^2) - (a_i u_i + b_i v_i + 2r_i \theta + 2r_i \omega_i) \} \\ & \quad - 2\rho|\theta| - 2\gamma \|\boldsymbol{\omega}\|_p \\ & \leq \sum_{i=1}^N t_i \{ \log(u_i v_i - (\theta + \omega_i)^2) - (a_i u_i + b_i v_i) - 2(r_i \omega_i + \gamma|\omega_i|) \} \\ & \quad - 2 \left(\sum_{i=1}^N t_i r_i \theta + \rho|\theta| \right), \end{aligned}$$

from the relationship $\sum_{i=1}^N t_i |\omega_i| \leq \|\boldsymbol{\omega}\|_\infty \leq \|\boldsymbol{\omega}\|_p$. Moreover, this upper-bound coincides with the original problem when $\boldsymbol{\omega} = \mathbf{0}_N$. Therefore, if

$\omega = \mathbf{0}_N$ is a maximizer of this upper-bound, it is also a maximizer of (15). From the derivative of the upper-bound over ω_i , we get that $\omega_i = 0$ is a maximizer if the following condition holds:

$$-(\gamma + r_i) \leq \frac{\theta}{u_i v_i - \theta^2} \leq (\gamma - r_i).$$

This is a sufficient condition for the original problem (15) to have $\omega_i = 0$ as its solution. Under this condition, problem (15) can be expressed as

$$\begin{aligned} \max_{\theta, \tilde{u}, \tilde{v}, u_i, v_i} \quad & \log(\tilde{u}\tilde{v} - \theta^2) - (\tilde{a}\tilde{u} + \tilde{b}\tilde{v}) - 2(\tilde{r}\theta + \rho|\theta|) \\ \text{s.t.} \quad & \tilde{u}\tilde{v} - \theta^2 > 0, \\ & -(\gamma + r_i) \leq \frac{\theta}{u_i v_i - \theta^2} \leq (\gamma - r_i) \quad (1 \leq i \leq N) \end{aligned}$$

for some properly chosen \tilde{a}, \tilde{b} and $\tilde{r} = \sum_{i=1}^N t_i r_i$. Hence, since the additional condition involves $\theta = 0$ irrelevant to the value of u_i and v_i if $\max_{1 \leq i \leq N} |r_i| \leq \gamma$ holds, we have $\theta = 0$ when $|\tilde{r}| \leq \rho$ from Idé et al. (2009, Proposition 1). \square

Appendix C.5. Proof of Theorem 3

Let $h(\mathbf{y}) = \frac{1}{2}\|\mathbf{y} - \mathbf{y}_0\|_2^2$ and \mathbf{y}' be one of the feasible solutions to the original problem (13). Moreover, since $\tilde{\mathbf{y}}$ is infeasible for the original problem (13), $\|\tilde{\mathbf{y}}\|_q > \gamma$ holds. Then, for $\mathbf{y}'' = \mathbf{y}' + \epsilon(\tilde{\mathbf{y}} - \mathbf{y}')$ with $0 < \epsilon \leq 1$, $h(\mathbf{y}'') \leq h(\mathbf{y}')$ holds from the convexity of h . Therefore, \mathbf{y}'' is a better solution to problem (13) as long as the constraints $|\mathbf{1}_N^\top \mathbf{y}''| \leq \rho$ and $\|\mathbf{y}''\|_q \leq \gamma$ are satisfied. The first condition always holds because $|\mathbf{1}_N^\top \mathbf{y}''| \leq (1 - \epsilon)|\mathbf{1}_N^\top \mathbf{y}'| + \epsilon|\mathbf{1}_N^\top \tilde{\mathbf{y}}| \leq \rho$. On the other hand, the latter condition $\|\mathbf{y}''\|_q = \left(\sum_{i=1}^N |y_i''|^q\right)^{\frac{1}{q}} \leq \gamma$ is no longer valid if $\|\mathbf{y}'\|_q = \gamma$ and $\text{sgn}(y_i') = \text{sgn}(\tilde{y}_i - y_i')$, which results in $\tilde{y}_i y_i' \geq 0$. This is a necessary condition for the solution to (13). Otherwise, we can always improve the solution by the above procedure, which contradicts its optimality. \square

References

Agarwal, A., Negahban, S., Wainwright, M., 2011. Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. Proceedings of the 28th International Conference on Machine Learning, 1129–1136.

- Ahmed, A., Xing, E. P., 2009. Recovering time-varying networks of dependencies in social and biological studies. *Proceedings of the National Academy of Sciences* 106 (29), 11878–11883.
- Bach, F. R., 2008. Consistency of the group lasso and multiple kernel learning. *The Journal of Machine Learning Research* 9, 1179–1225.
- Baillie, R. T., Bollerslev, T., 1989. Common stochastic trends in a system of exchange rates. *The Journal of Finance* 44 (1), 167–181.
- Banerjee, O., El Ghaoui, L., d’Aspremont, A., 2008. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *The Journal of Machine Learning Research* 9, 485–516.
- Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J., 2011. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning* 3 (1), 1–122.
- Boyd, S., Vandenberghe, L., 2004. *Convex optimization*. Cambridge University Press.
- Candès, E. J., Li, X., Ma, Y., Wright, J., 2011. Robust principal component analysis? *Journal of the ACM* 58 (3), 11:1–11:37.
- Caruana, R., 1997. Multitask learning. *Machine Learning* 28 (1), 41–75.
- Chandrasekaran, V., Parrilo, P., Willsky, A., 2010. Latent variable graphical model selection via convex optimization. *Arxiv preprint arXiv:1008.1290*.
- Chiquet, J., Grandvalet, Y., Ambroise, C., 2011. Inferring multiple graphical structures. *Statistics and Computing* 21 (4), 537–553.
- Dempster, A. P., 1972. Covariance selection. *Biometrics* 28 (1), 157–175.
- Duchi, J., Gould, S., Koller, D., 2008a. Projected subgradient methods for learning sparse gaussians. *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence*, 145–152.
- Duchi, J., Shalev-Shwartz, S., Singer, Y., Chandra, T., 2008b. Efficient projections onto the l_1 -ball for learning in high dimensions. *Proceedings of the 25th international conference on Machine learning*, 272–279.

- Durbin, J., Koopman, S., Atkinson, A., 2001. Time series analysis by state space methods. Vol. 15. Oxford University Press.
- Fan, J., Feng, Y., Wu, Y., 2009. Network exploration via the adaptive lasso and scad penalties. *The Annals of Applied Statistics* 3 (2), 521.
- Friedman, J., Hastie, T., Tibshirani, R., 2008. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* 9 (3), 432–441.
- Gabay, D., Mercier, B., 1976. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers & Mathematics with Applications* 2 (1), 17–40.
- Golub, G., Van Loan, C., 1996. Matrix computations. Vol. 3. Johns Hopkins University Press.
- Guo, F., Hanneke, S., Fu, W., Xing, E., 2007. Recovering temporally rewiring networks: A model-based approach. In: *Proceedings of the 24th International Conference on Machine learning*. ACM, pp. 321–328.
- Guo, J., Levina, E., Michailidis, G., Zhu, J., 2011. Joint estimation of multiple graphical models. *Biometrika* 98 (1), 1–15.
- Hamilton, J., 1994. Time series analysis. Vol. 2. Cambridge University Press.
- Hara, S., Kawahara, Y., Washio, T., von Büna, P., Tokunaga, T., Yumoto, K., 2012. Separation of stationary and non-stationary sources with a generalized eigenvalue problem. *Neural Networks* 33, 7–20.
- Hara, S., Washio, T., 2011. Common substructure learning of multiple graphical gaussian models. *Machine Learning and Knowledge Discovery in Databases*, 1–16.
- Hestenes, M., 1969. Multiplier and gradient methods. *Journal of Optimization Theory and Applications* 4 (5), 303–320.
- Honorio, J., 2011. Lipschitz parametrization of probabilistic graphical models. *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence*, 347–354.

- Honorio, J., Samaras, D., 2010. Multi-task learning of gaussian graphical models. *Proceedings of the 27th International Conference on Machine Learning*, 447–454.
- Hsieh, C., Sustik, M., Dhillon, I., Ravikumar, P., 2011. Sparse inverse covariance matrix estimation using quadratic approximation. *Advances in Neural Information Processing Systems* 24, 2330–2338.
- Idé, T., Lozano, A. C., Abe, N., Liu, Y., 2009. Proximity-based anomaly detection using sparse structure learning. *Proceedings of the 2009 SIAM International Conference on Data Mining*, 97–108.
- Jalali, A., Ravikumar, P., Sanghavi, S., Ruan, C., 2010. A dirty model for multi-task learning. *Advances in Neural Information Processing Systems* 23, 964–972.
- Lauritzen, S., 1996. *Graphical models*. Oxford University Press, USA.
- Li, L., Toh, K., 2010. An inexact interior point method for l_1 -regularized sparse covariance selection. *Mathematical Programming Computation*, 1–25.
- Liu, H., Palatucci, M., Zhang, J., 2009. Blockwise coordinate descent procedures for the multi-task lasso, with applications to neural semantic basis discovery. *Proceedings of the 26th International Conference on Machine Learning*, 649–656.
- Meinshausen, N., Bühlmann, P., 2006. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics* 34 (3), 1436–1462.
- Obozinski, G., Jacob, L., Vert, J., 2011. Group lasso with overlaps: the latent group lasso approach. *Arxiv preprint arXiv:1110.0413*.
- Powell, M., 1967. A method for non-linear constraints in minimization problems. *Optimization*, 283–298.
- Scheinberg, K., Ma, S., Goldfarb, D., 2010. Sparse inverse covariance selection via alternating linearization methods. *Advances in Neural Information Processing Systems* 23, 2101–2109.

- Scheinberg, K., Rish, I., 2010. Learning sparse gaussian markov networks using a greedy coordinate ascent approach. *Machine Learning and Knowledge Discovery in Databases*, 196–212.
- Sra, S., 2011. Fast projections onto $\ell_{1,q}$ -norm balls for grouped feature selection. *Machine Learning and Knowledge Discovery in Databases*, 305–317.
- Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B* 58 (1), 267–288.
- Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., Knight, K., 2005. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B* 67 (1), 91–108.
- Tomioka, R., Suzuki, T., Sugiyama, M., 2011. Super-linear convergence of dual augmented lagrangian algorithm for sparsity regularized estimation. *The Journal of Machine Learning Research* 12, 1537–1586.
- Turlach, B., Venables, W., Wright, S., 2005. Simultaneous variable selection. *Technometrics* 47 (3), 349–363.
- Varoquaux, G., Gramfort, A., Poline, J. B., Thirion, B., 2010. Brain covariance selection: better individual functional connectivity models using population prior. *Advances in Neural Information Processing Systems* 23, 2334–2342.
- von Büna, P., Meinecke, F. C., Király, F. C., Müller, K. R., 2009. Finding stationary subspaces in multivariate time series. *Physical Review Letters* 103 (21), 214101.
- Wainwright, M., Ravikumar, P., Lafferty, J., 2007. High-dimensional graphical model selection using ℓ_1 -regularized logistic regression. *Advances in Neural Information Processing Systems* 19, 1465–1472.
- Wang, C., Sun, D., Toh, K., 2009. Solving log-determinant optimization problems by a newton-cg primal proximal point algorithm. *SIAM Journal on Optimization* 20, 2994–3013.
- Yuan, M., Lin, Y., 2006. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B* 68 (1), 49–67.

- Yuan, M., Lin, Y., 2007. Model selection and estimation in the gaussian graphical model. *Biometrika* 94, 19–35.
- Yuan, X., 2009. Alternating direction methods for sparse covariance selection. Preprint available at http://www.optimization-online.org/DB_HTML/2009/09/2390.html.
- Zhang, B., Li, H., Riggins, R. B., Zhan, M., Xuan, J., Zhang, Z., Hoffman, E. P., Clarke, R., Wang, Y., 2009. Differential dependency network analysis to identify condition-specific topological changes in biological networks. *Bioinformatics* 25 (4), 526–532.
- Zhang, B., Wang, Y., 2010. Learning structural changes of gaussian graphical models in controlled experiments. *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence*, 701–708.
- Zhou, S., Lafferty, J., Wasserman, L., 2010. Time varying undirected graphs. *Machine Learning* 80 (2), 295–319.
- Zou, H., 2006. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* 101 (476), 1418–1429.

Table 3: Simulation results for three cases ($d = 25, 50$ and 100) with $N = 5$ datasets evaluated by weighted precision, recall and F-measure, denoted by "Prec.", "Rec." and "F" in the table, respectively. The " F_0 " denotes the F_0 -measure for zero pattern identification. Each simulation is conducted so that each dataset has $5d$ data points, and the measurements are averaged over 100 random realization of datasets. The numbers in brackets are standard deviations of each measurement. Each of the three rows in SICS and MSICS corresponds to results for $\epsilon_0 = 0.5, 0.7$ and 0.9 from the top. We highlight the top three results for each measurement in bold font (except for " F_0 ").

		CSSL ($p = 1$)	CSSL ($p = 2$)	CSSL ($p = \infty$)	CSSL ($\gamma = \infty$)	SICS	MSICS ($p = 2$)	MSICS ($p = \infty$)
$d = 25$	Prec.	.84 (.19)	.70 (.16)	.56 (.19)	.48 (.20)	.14 (.14)	.38 (.21)	.54 (.23)
						.20 (.16)	.43 (.21)	.49 (.21)
						.33 (.16)	.41 (.19)	.45 (.19)
	Rec.	.45 (.32)	.82 (.14)	.84 (.12)	.86 (.11)	.07 (.07)	.48 (.24)	.60 (.24)
						.23 (.18)	.74 (.19)	.74 (.19)
						.80 (.20)	.83 (.13)	.86 (.11)
	F	.56 (.22)	.75 (.14)	.66 (.17)	.60 (.19)	.09 (.08)	.41 (.21)	.55 (.23)
						.21 (.16)	.53 (.21)	.58 (.20)
						.45 (.18)	.53 (.19)	.58 (.18)
	F_0	.92 (.02)	.92 (.02)	.92 (.02)	.92 (.02)	.92 (.02)	.93 (.02)	.92 (.02)
$d = 50$	Prec.	.87 (.11)	.69 (.14)	.56 (.17)	.47 (.17)	.10 (.13)	.24 (.20)	.58 (.19)
						.13 (.14)	.37 (.20)	.52 (.19)
						.27 (.19)	.42 (.18)	.47 (.18)
	Rec.	.41 (.20)	.83 (.11)	.85 (.10)	.91 (.05)	.04 (.04)	.18 (.19)	.60 (.19)
						.10 (.11)	.51 (.21)	.72 (.16)
						.50 (.22)	.81 (.12)	.86 (.08)
	F	.53 (.20)	.75 (.12)	.66 (.15)	.61 (.15)	.05 (.06)	.20 (.19)	.58 (.19)
						.10 (.11)	.42 (.20)	.59 (.18)
						.34 (.20)	.54 (.17)	.60 (.16)
	F_0	.90 (.03)	.90 (.02)	.89 (.02)	.89 (.03)	.89 (.03)	.90 (.02)	.90 (.03)
$d = 100$	Prec.	.91 (.07)	.78 (.10)	.64 (.14)	.53 (.15)	.09 (.11)	.17 (.14)	.68 (.15)
						.10 (.12)	.33 (.21)	.62 (.16)
						.22 (.17)	.46 (.18)	.55 (.16)
	Rec.	.37 (.18)	.81 (.11)	.83 (.11)	.95 (.02)	.03 (.10)	.06 (.10)	.59 (.17)
						.06 (.10)	.25 (.21)	.67 (.15)
						.24 (.19)	.67 (.16)	.82 (.09)
	F	.51 (.19)	.79 (.10)	.72 (.12)	.67 (.12)	.05 (.10)	.08 (.11)	.63 (.16)
						.07 (.10)	.28 (.21)	.64 (.15)
						.22 (.18)	.54 (.17)	.65 (.14)
	F_0	.87 (.04)	.87 (.04)	.87 (.03)	.87 (.03)	.87 (.03)	.88 (.04)	.87 (.03)

Table 4: Anomaly detection results: The simulation is conducted for 4 different settings, $[n_n, n_f] = [4, 1], [12, 3], [20, 5]$ and $[40, 10]$. For each method, we compute precision matrices for 11 different values of α ranging from $10^{-1.5}$ to $10^{-0.5}$. The table shows the median of the best AUCs among these 11 results over 100 random realizations of datasets. The numbers in brackets are 25% and 75% quantiles. The bold font represents the top three results, which are CSSL ($p = 2$), CSSL ($p = \infty$) and MSICS ($p = \infty$) for all settings.

	$[n_n, n_f] = [4, 1]$		$[n_n, n_f] = [12, 3]$	
	best AUC	α	best AUC	α
CSSL ($p = 1$)	.975 (.950 / .987)	$10^{-0.9}$.975 (.950 / 1.00)	$10^{-0.9}$
CSSL ($p = 2$)	.987 (.963 / 1.00)	$10^{-0.9}$.987 (.963 / 1.00)	$10^{-0.9}$
CSSL ($p = \infty$)	.987 (.963 / 1.00)	$10^{-0.9}$	1.00 (.987 / 1.00)	$10^{-0.9}$
SICS	.975 (.938 / .987)	$10^{-0.5}$.975 (.938 / .987)	$10^{-0.5}$
MSICS ($p = 2$)	.975 (.950 / .987)	$10^{-0.8}$.975 (.950 / .987)	$10^{-0.7}$
MSICS ($p = \infty$)	.987 (.963 / 1.00)	$10^{-1.1}$.987 (.975 / 1.00)	$10^{-1.2}$
	$[n_n, n_f] = [20, 5]$		$[n_n, n_f] = [40, 10]$	
	best AUC	α	best AUC	α
CSSL ($p = 1$)	.975 (.950 / 1.00)	$10^{-0.9}$.975 (.963 / 1.00)	$10^{-0.9}$
CSSL ($p = 2$)	1.00 (.975 / 1.00)	$10^{-0.8}$.987 (.963 / 1.00)	$10^{-0.8}$
CSSL ($p = \infty$)	1.00 (.987 / 1.00)	$10^{-0.9}$	1.00 (.987 / 1.00)	$10^{-0.9}$
SICS	.975 (.950 / .987)	$10^{-0.5}$.975 (.950 / .987)	$10^{-0.5}$
MSICS ($p = 2$)	.975 (.950 / .987)	$10^{-1.0}$.975 (.950 / .987)	$10^{-1.0}$
MSICS ($p = \infty$)	.987 (.975 / 1.00)	$10^{-1.1}$.987 (.975 / 1.00)	$10^{-0.9}$